January 16, 2006

Battelle Validation Sampling Design Software – Beta Release

User's Guide

by

BATTELLE
505 King Avenue
Columbus, OH 43201-2693

Contract No. 282-98-0019
Task 16

Dr. Peter Scheidt
Project Officer

Dr. Jim Quackenboss
Dr. Warren Galke
Work Assignment Officers

National Children's Study Program Office
National Institute for Child Health and Human Development
6100 Executive Blvd - 5C01
Rockville, MD 20892

TABLE OF CONTENTS

TABLE OF CONTENTS

Page

List of Tables

List of Figures

1 Installation

Recommended computer hardware and operating system requirements are:

Operating System:	Microsoft® Windows XP or Windows 2000
Computer Hardware:	256 MB RAM
	50 MB of available disk space
	1.5 GHz processor speed

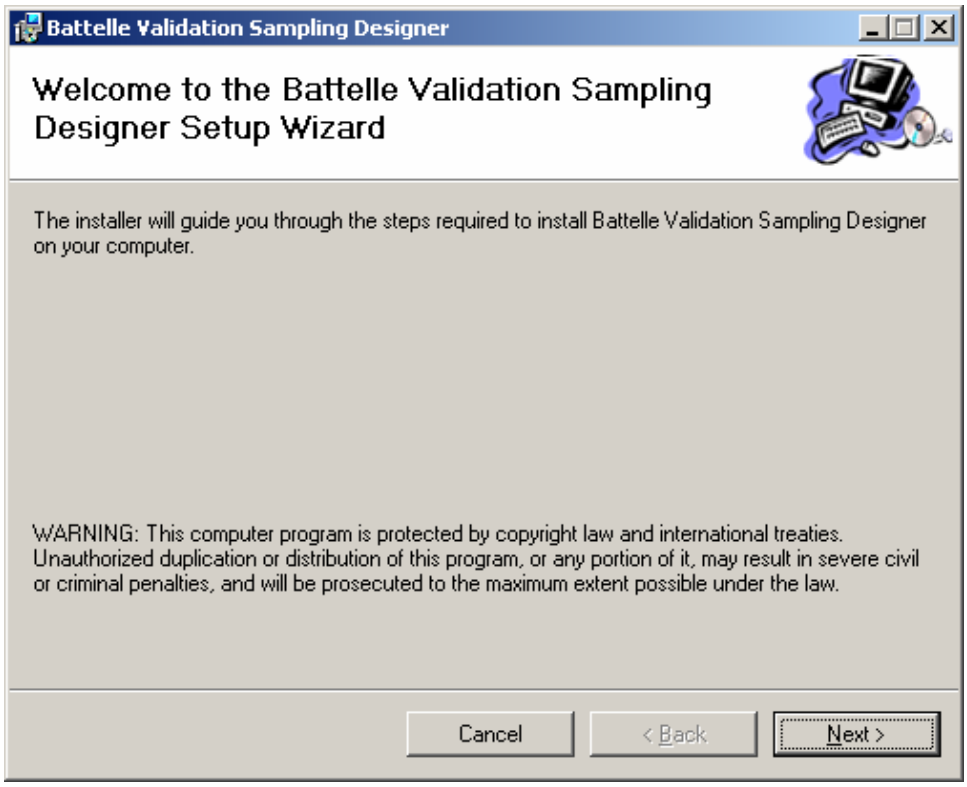
It is also recommended that the user chooses a Windows Display Setting with a screen resolution of 1280 by 1024 pixels, or greater.

The following three files are required for installation of the Battelle Validation Sampling Designer Software (Beta Release) on to a personal computer.

1. Setup.exe
2. Setup.msi
3. Setup.ini

To install the software, copy the above three files into the same directory and launch the program Setup.exe – which will guide the user through the set-up process:

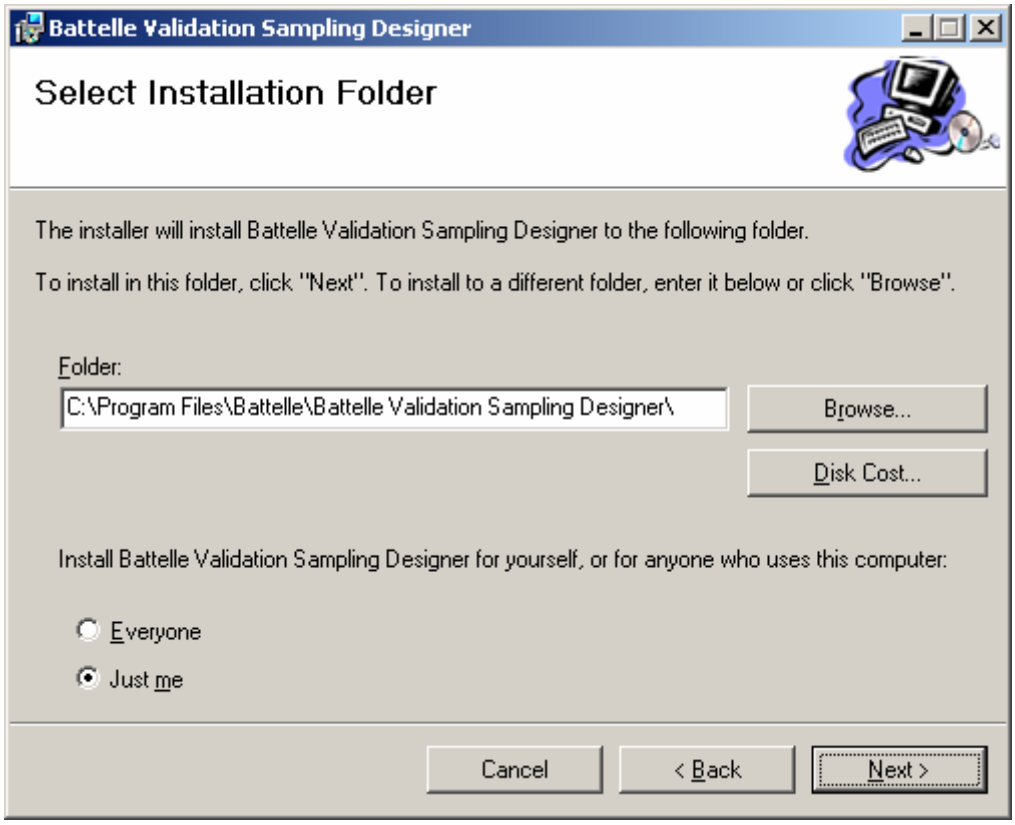
- The first screen in the set-up provides a welcome message and copyright information. Click the 'Next' button to continue with the installation, or 'Cancel' to exit the installation.



- The second screen allows the user to select the folder on the computer where the software will be installed. The default location for the software is "C:\Program Files\Battelle\Battelle Validation Sampling Designer", however the user can select a different location if desired by using the browse button or typing in an alternative location.

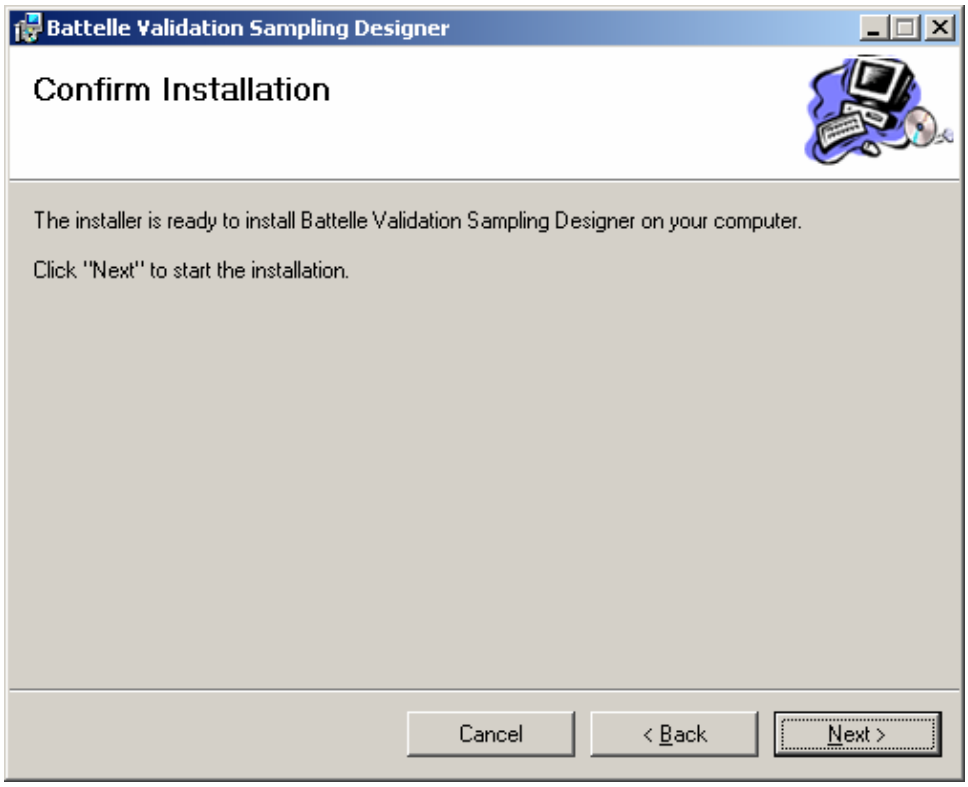
    - There is a button labeled 'Disk Cost' which will show the user how much disk space is available on the computer for the software.

    - For operating systems that allow for multiple users to log onto the same computer, there is a radio button that allows access to either (1) the current user only, or (2) all users on this computer

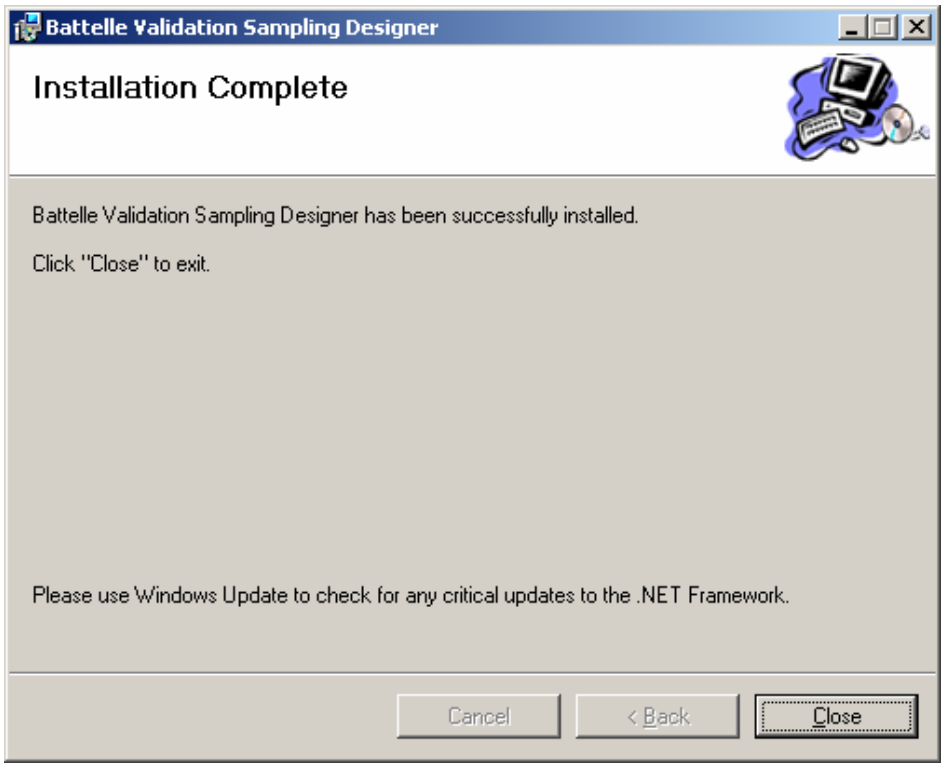


- The third screen allows the user to confirm the installation options. Click the 'Next' button to proceed, the 'Back' button to navigate to the previous screen, or the 'Cancel' button to exit the installation.

    - After clicking the 'Next' button, the software will be installed onto the computer. Typical installation time is less than one minute.



- The last screen provides confirmation of successful installation and instructs the user to click ‘Close’ to exit the installation process.



Once the software is successfully installed, the user can access the Battelle Validation Sampling Designer through the Programs directory that is accessible from the Windows® 'Start' button, under the heading Battelle/Validation Sampling Designer.

2 Disclaimer and Restricted Rights

The Battelle Validation Sampling Designer computer software is a work prepared by Battelle under Government Contract No. 282-98-0019, Task Order 11, Work Assignment 16 for the National Children's Study Program Office at the National Institute of Child Health and Human Development (NICHD). The software represents a Beta-Version of a statistical sampling design tool developed to allow planners of the National Children's Study investigate hypothesis-specific study designs that incorporate validation sampling approaches. As such, this software product is being released to NICHD as an executable under the terms and conditions set forth below in the Restricted Rights Notice, and is intended for use by the Government Agencies (NICHD, NIEHS, CDC and EPA) for the express purpose of planning the National Children's Study.

This software and any data it generates is copyright protected. In no event shall either the National Institute of Child Health and Human Development (NICHD) or Battelle have any responsibility or liability for any consequences of any use, misuse, inability to use, or reliance upon the information contained within or generated by the software. Nor does either the NICHD, or Battelle, warrant or otherwise represent in any way the accuracy, adequacy, efficacy, or applicability of the software or data generated by the software.

Restricted Rights Notice (Jun 1987)

(a) This computer software is submitted with restricted rights under Government Contract No.282-98-0019, Task Order 11, Work Assignment 16. It may not be used, reproduced, or disclosed by the Government except as provided in paragraph (b) of this Notice or as otherwise expressly stated in the contract.

(b) This computer software may be--

(1) Used or copied for use in or with the computer or computers for which it was acquired, including use at any Government installation to which such computer or computers may be transferred;

(2) Used or copied for use in a backup computer if any computer for which it was acquired is inoperative;

(3) Reproduced for safekeeping (archives) or backup purposes;

(4) Modified, adapted, or combined with other computer software, *provided* that the modified, combined, or adapted portions of the derivative software incorporating restricted computer software are made subject to the same restricted rights;

(5) Disclosed to and reproduced for use by support service contractors in accordance with subparagraphs (b)(1) through (4) of this clause, provided the Government makes such disclosure or reproduction subject to these restricted rights; and

(6) Used or copied for use in, or transferred to, a replacement computer.

(c) Notwithstanding the foregoing, if this computer software is published, copyrighted computer software, it is licensed to the Government, without disclosure prohibitions, with the minimum rights set forth in paragraph (b) of this clause.

(d) Any other rights or limitations regarding the use, duplication, or disclosure of this computer software are to be expressly stated in, or incorporated in, the contract.

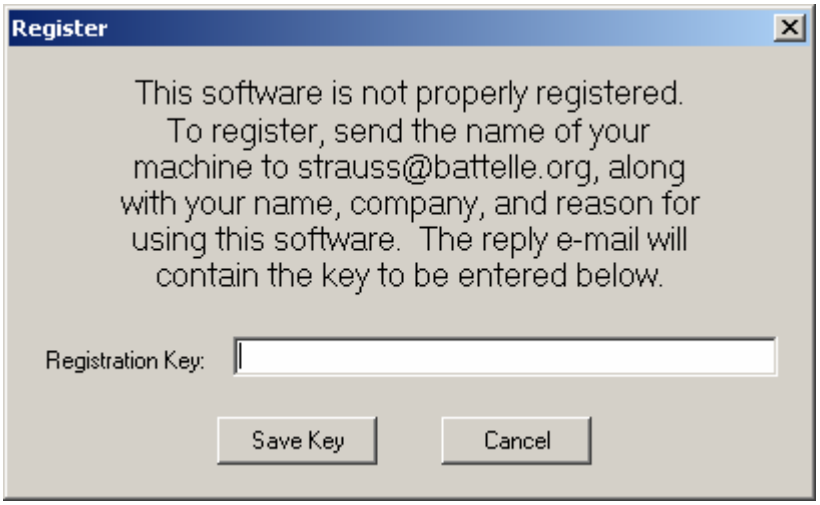
(e) This Notice shall be marked on any reproduction of this computer software, in whole or in part.

3 Registration

As explained above in Section 2 of the Users Guide, this prototype software is being released as an executable, and is intended for use by the Government Agencies (NICHD, NIEHS, CDC and EPA) for the express purpose of planning the National Children's Study. As such, the executable requires the user to register the software product with Battelle prior to use, and will not allow the user to utilize the software without completing the registration process. The registration process is designed to facilitate two functions:

- It will limit the distribution of the prototype software tool only to authorized users that are working on the National Children's Study; and
- It will allow Battelle to maintain a list of authorized users so that future releases of the software can be efficiently distributed as they become available.

Upon using the Battelle Validation Sampling Designer for the first time, a registration window will appear that asks the user to register the software with Battelle by sending an e-mail to strauss@battelle.org with information about the user and the specific computer that the software was installed upon.



The information contained in this e-mail should include the following:

- Name of the user
- Contact information for the user
  - Name of Company/Agency
  - Mailing address
  - Phone number
  - e-mail address
- Computer name
- A short description of the user's reason for using this software

To obtain the name of your computer, right-click on the 'My Computer' icon on your desktop, select (left-click) 'Properties', and select the 'Computer Name' tab.

Once Battelle receives the e-mail with the above information, we will (1) verify the user's role in the National Children's Study with NICHD; and (2) reply to the e-mail with a text key that will need to be copied and pasted by the user into the Registration Key text window.

Please note that the Registration Key and the algorithm used to generate the Registration Key (based on the computer name) are case-sensitive.

Once the Registration Key has been successfully entered, the software will engage and the Battelle Validation Sampling Designer will be launched with the first screen displaying the Disclaimer and Restricted Rights information contained in Section 2 of this document. The Registration Key is computer specific, and will only allow the software to be launched from the specific computer under which the software was registered.

4 General Description of the Software Tool

As part of work in support of the National Children's Study, Battelle has collaborated with faculty from the Department of Biostatistics at Harvard University to develop innovative (and cost saving) statistical approaches for gathering exposure information for this study that focus on assessing relationships between adverse health outcomes and complex environmental exposures. The statistical techniques that have been developed concentrate on using a validation sampling approach – in which detailed exposure assessment information is collected on a carefully selected subset of the study population, while cheaper (and presumably less accurate and less burdensome) surrogate measures are collected across the entire cohort. Preliminary work conducted by Battelle and Harvard indicated the potential for large scale efficiencies in sampling and analysis costs (both dollars and study subject burden) associated with the validation sampling approach compared to traditional study designs in which detailed exposure information is collected on all study participants. These efficiency gains can be realized without making any substantive sacrifices on the ability to draw unbiased inferences on the research problem of characterizing the relationship between exposure and health outcome.

The conceptual sampling approaches that have been developed allow for multiple stages of sampling for exposure information - where stages represent increasing levels of information, complexity and cost with respect to exposure assessment (e.g., from simple questionnaire information, to biomarker data, to aggregate exposure assessment information, to repeated waves of aggregate exposure information). The goal of the statistical sampling approaches is to determine what fraction (and specific members) of a study population should be assessed at each stage to allow the study planner to determine the most cost efficient (or burden efficient) manner to design the study so that the relationship between disease and exposure can be appropriately characterized.

Additional efficiencies can be gained in the sampling strategy by allowing for outcome dependent or covariate dependent designs (in which the decision to sample exposure for a particular study subject is dependent on either the health outcome (e.g., analysis of archived samples), or on previously collected lower-stage exposure information).

The framework for the validation sampling approach relies on the implementation of staged sampling, in which Stage 0 represents the fraction of the cohort on which health outcome (*and effect modifier*) information is collected (denoted Y *and E*), Stage 1 represents a subset of the Stage 0 participants who are measured for the lowest-level surrogate measure (denoted Z_1), Stage 2 represents a subset of the Stage 1 participants who are measured for the next level surrogate measure (denoted Z_2), and so on. The last stage of sampling corresponds to measuring the true exposure of interest (denoted X). In the case in which X is not measurable, the last stage of sampling will include not participants and a latent variable approach is pursued.

It should be noted that the stages of sampling in the design do not correspond to a temporal ordering for when the samples are collected – in fact the health outcome which is measured in the first stage of sampling (Stage 0) will likely be measured after all other samples (exposure variables and surrogate measures of exposure) are collected. Rather, the stages correspond to an assumed hierarchy of sampling, in which Stage 0 represents a sub-sample of study subjects in the NCS cohort that will have the health outcome measured, Stage 1 represents a subset of participants who were selected in Stage 0 who have a surrogate measure of exposure, and so on.

For a two-stage surrogate measure design in which the true exposure is measurable, the software is capable of developing an optimal design as suggested by the following series of logistic regression equations:

$$\text{logit}(\gamma_0) = \alpha_{00}$$
$$\text{logit}(\gamma_1) = \alpha_{10} + \alpha_{11} \cdot Y$$
$$\text{logit}(\gamma_2) = \alpha_{20} + \alpha_{21} \cdot Y + \alpha_{22} \cdot Z_1$$
$$\text{logit}(\gamma_3) = \alpha_{30} + \alpha_{31} \cdot Y + \alpha_{32} \cdot Z_1 + \alpha_{33} \cdot Z_2, \text{ where}$$

- γ_0 represents the probability that a study subject is sampled at stage 0 for the health outcome (Y),
- γ_1 represents the conditional probability that a study subject is sampled at stage 1 (for the surrogate measure Z_1), given that (s)he was already selected within stage 0,
- γ_2 represents the conditional probability that a study subject is sampled at stage 2 (for the surrogate measure Z_2), given that (s)he was already selected within stage 1, and
- γ_3 represents the conditional probability that a study subject is sampled at stage 3 (for the exposure variable X), given that (s)he was already selected within stage 2.

The intercept terms (α_{00}, α_{10}, α_{20}, and α_{30}) in the above series of logistic regression models are included in all designs with two stages of surrogate measures and a measurable exposure variable. The slope terms (α_{11}, α_{21}, and α_{31}) represent parameters associated with outcome dependent sampling in stages 1, 2, and 3; and the slope terms

(α_{22}, α_{32}, and α_{33}) represent parameters associated with covariate dependent sampling in stages 2 and 3. Thus, a design with no opportunity for covariate dependent or outcome dependent sampling would constrain these slope (α) terms to zero. A design that includes random sampling for the surrogate measures (Z_1 and Z_2) and outcome dependent sampling for the exposure variable (X) would include all four intercept terms (α_{00}, α_{10}, α_{20}, and α_{30}) and the slope term α_{31}.

Due to the hierarchical nature of the staged sampling approach, the actual probability of sampling at any stage is a product of the γ probabilities from the current and all previous stages of sampling. Thus, the overall probability of being sampled for X in stage 3 (as a function of α,Y, Z_1 and Z_2) is $\pi_3 = \gamma_0 \cdot \gamma_1 \cdot \gamma_2 \cdot \gamma_3$.

Our approach for optimal design of validation sampling design is based on likelihood methods, in which critical inputs are specified by the user (distribution of X, Y|X, Z_i|X for all i; costs associated with measuring Y, X, Z_i; the α constraints that dictate the possibilities of covariate or outcome dependent sampling at each stage; and γ constraints that allow us to collapse two stages of sampling together – so that, for example, Y and Z_1 are measured together on the same fraction of the cohort), and then the optimal design (dictated by the α's in the above series of logistic regression models) is identified using numeric constrained optimization. Our approach allows for two general types of constrained optimization: (1) the user can specify a target total dollar budget to be spent on the data collection effort, and the constrained optimization will determine the set of α's that result in a design with the lowest possible standard error for the parameter of interest, or (2) the user can specify a target standard error for the parameter of interest, and the constrained optimization will determine the set of α's that result in a design with the lowest possible total dollar budget to be spent on the data collection effort.

The current (Beta-Release) software tool allows flexibility in terms of the research problems that it will support, including: the inclusion of effect modifiers in the model between exposure and health outcome, multiple distributional assumptions for the outcome and exposure variables, the ability for the exposure variable to be either measurable or latent, and different options for the constrained optimization in the design. A subsequent planned version of this software will also allow the user to explore a limited number of designs that include repeated measures for the health outcome data in a longitudinal model.

The Battelle Validation Sampling Designer software collects information employing a structured interview that allows the user to input appropriate statistical details about the specific hypothesis that is being investigated. A summary of the information gathered in the structured interview is provided in the bulleted list below:

- High-Level Summary of the Design
    - Number of stages of surrogate measures to be considered
    - Maximum sample size of the cohort (assumed to be 100,000 for the National Children's Study)
    - Presence/Absence of an effect modifier

- Health outcome is represented by a cross-sectional measure or a series of repeated measures in the design. (*Note that in the Beta-release, only the cross-sectional designs are supported by the software*).

- Marginal Distribution (and Cost) of the Exposure Variable
  - User chooses among three potential distributions (Normal, Lognormal, or Binomial) and provides values for the parameters that define the marginal distribution of exposure.
  - User chooses whether the Exposure Variable is measurable
    - If the Exposure Variable is measurable, then the user is asked to provide information related to the costs of exposure assessment. Cost input is split into sampling and analytical costs, with analytical costs assumed to be associated with the last stage of sampling and sampling costs allowed to be allocated to the last stage of sampling or any previous stage of sampling.
    - If the Exposure Variable is not measurable, a latent variable design is assumed – and the user will need to input a design with a minimum of two stages of surrogate measures.

- Distribution and Cost of the Effect Modifier
  - This information is gathered only if the user selects to include an Effect Modifier in the High-Level Summary of the Design. It is automatically assumed that the effect modifier is assessed at a stage in-between the health outcome (stage 0) and the lowest-level surrogate measure (stage 1).
  - User chooses among three potential distributions (Normal, Lognormal, or Binomial) and provides values for the parameters that define the marginal distribution of the Effect Modifier.
  - User provides information on the statistical relationship between the Exposure Variable and the Effect Modifier
    - A simple check-box is included to ascertain whether they are independent
    - If there is a dependency between the Exposure Variable and the Effect Modifier, the software allows the user to input statistical information that defines this relationship.
  - User provides information related to the costs of assessing the value of the effect modifier. Cost input is split into sampling and analytical costs, with analytical costs tied to the same stage as the effect modifier and sampling costs allowed to be allocated to the effect modifier stage or the previous stage that includes the health outcome.

- Distribution of the Health Outcome, as a Function of Exposure (and Effect Modifier) and Associated Cost
  - User chooses among three potential distributions (Normal, Lognormal, or Binomial) that define the conditional distribution of the Health Outcome, given Exposure (and Effect Modifier).

- User provides information on the statistical relationship between the Health Outcome and the Exposure Variable (and Effect Modifier)
- User provides information related to the costs of assessing the value of the Health Outcome.

- Distribution and Cost of (multiple) Surrogate Measures of Exposure
  - User defines the stage of sampling associated with the Surrogate Measure
  - User chooses among three potential distributions (Normal, Lognormal, or Binomial) that define the conditional distribution of the Surrogate Measure, given true Exposure.
  - User provides information on the statistical relationship between the Surrogate Measure and the Exposure Variable
    - In cases in which both the exposure variable and the surrogate measure are continuous, the statistical relationship can be assumed to follow the 'Classic Measurement Error Model' in which Z=X+Error. When both variables are continuous, the user can also explore relationships that assume some additive and/or multiplicative bias, in which case $Z=\beta_0 + \beta_1 \cdot X + \text{Error}$ – where $\beta_0 \neq 0$ and/or $\beta_1 \neq 1$.
    - In the event that the user specifies the exposure variable (X) as latent (not measurable), either (1) two stages of surrogate sampling must be employed in which on of those stages includes a surrogate measure where the relationship with the exposure variable is defined by the 'Classic Measurement Error Model', or (2) three (or more) stages of surrogate sampling must be employed.
  - User provides information related to the costs of assessing the Surrogate Measure. Cost input is split into sampling and analytical costs, with analytical costs assumed to be associated with the stage of sampling identified for the surrogate measure and sampling costs allowed to be allocated to any stage of sampling that is less than or equal to the stage of sampling identified for the surrogate measure.
  - The above information is gathered for as many potential surrogate measures as the user would like to investigate. The user can input multiple candidate surrogate measures within the same stage of sampling – which will result in the software finding a unique solution for each possible combination of surrogate measures.

- Constrained Optimization Goals and Parameters
  - The software allows the user to either:
    - Constrain the standard error associated with the parameter of interest, and utilize the validation sampling designer software to determine the lowest-cost sampling plan.
    - Constrain the total budget associated with the study design, and utilize the validation sampling designer software to determine the design that leads to the smallest standard error for the parameter of interest.

- The software allows the user to input multiple constrained optimization goals – with each constraint optimized individually by the software.
- The software allows the user to compute the standard error constraint as a function of statistical size, power, and determination of a one-sided vs two-sided test.
- The software also allows the user to constrain sampling for each potential variable to the same subset as was sampled in the previous stage. As an example, this will allow the user to dictate that the Health Outcome and the lowest-level surrogate measure are measured on the same subset of study participants.

- Covariate and Outcome Dependent Sampling
  - The software includes a check-box lower triangular matrix that allows the user to input whether outcome or covariate dependent sampling should be pursued for each variable in Stages 1 and above in the design.

- Parameters that Govern the Numerical Optimization of the Design
  - The user can input
    - The number of integrations per subdivision
    - The tolerance at which the optimization will halt iteration
    - The Lagrange Multiplier that is used to impose the standard error or budget constraint
    - Scale values for each variable in the design (software only allows user to enter this information if a single combination of variables at each stage is explored under a single optimization constraint)
    - User designated or computer generated random starting values for the alpha parameters in the design (software only allows user to enter this information if a single combination of variables at each stage is explored under a single optimization constraint)

Once the user completes the structured interview, the software provides an html file description of the input that was provided for the design optimization, allowing the user to review the input prior to submitting the design scenario to the analytic optimizer.

Since the software allows the user to input multiple candidate surrogate measures at each stage, and multiple constrained optimization goals – the software is designed to explore each combination of potential designs specified by the user. This feature was added to allow the user to quickly screen multiple candidate surrogate measures and/or multiple types of constraints under a single submission to the analytic optimizer. However, there are many different parameters (e.g. Lagrange Multiplier, Tolerance) that govern the manner in which the software finds an appropriate optimal design – and the specific values of these parameters that lead to an optimal design can often be different for each combination of surrogate measures and constraints being explored. We therefore recommend that following the completion of a preliminary screening among multiple potential constraints/surrogate measures simultaneously – that the user select a single constraint and single surrogate measures at each stage to more carefully refine the

optimization (see section 5.9 on tips for finding the optimal solution using the Lagrange Multiplier).

Once the design input is submitted to the analytic optimizer, a screen will appear that will show progress while the optimizer is engaged with higher-level summary information on the design solution that is found for each combination of design specifications. Once the analytical optimizer is finished running, the user can use this screen to explore and sort the solutions (by clicking on the variable headers) prior to saving the results.

The saved results provide a summary of all designs explored in the submission. For each design, the software provides:

- A summary of the numeric optimization parameters
- A summary of the variable included at each stage of sampling, including
  - Costs
  - Dependencies on other variables included in the design
- A summary of how well the constraint was satisfied, and the optimum solution
  - For example, in a design specification in which the user constrains the standard error of the parameter of interest, the output provides a measure of the standard error achieved by the design – as well as the cost associated with implementing the design.
- The α parameters associated with the optimal design.

5 The Structured Interview

The following sections provide details regarding the structured interview in which the user provides input on the statistical design that is being optimized using the validation sampling approach. These sections provide pictures of the Graphical User Interface, as well as text descriptions of how to use the software to implement different useful design options with this software tool.

5.1 The Welcome Window

Following successful installation and registration of the Battelle Validation Sampling Designer, the software product will launch with a window that displays the Disclaimer and Restricted Rights Notice that is provided in Section 2 of this document. This window (depicted in Figure 1) will be displayed each time the software product is launched.

The scroll bar can be used to navigate to see the full Disclaimer and Restricted Rights Notice. To exit the software without proceeding to use the tool, click on the 'Cancel' button. To proceed with use of the software tool, click on the 'Next' button.

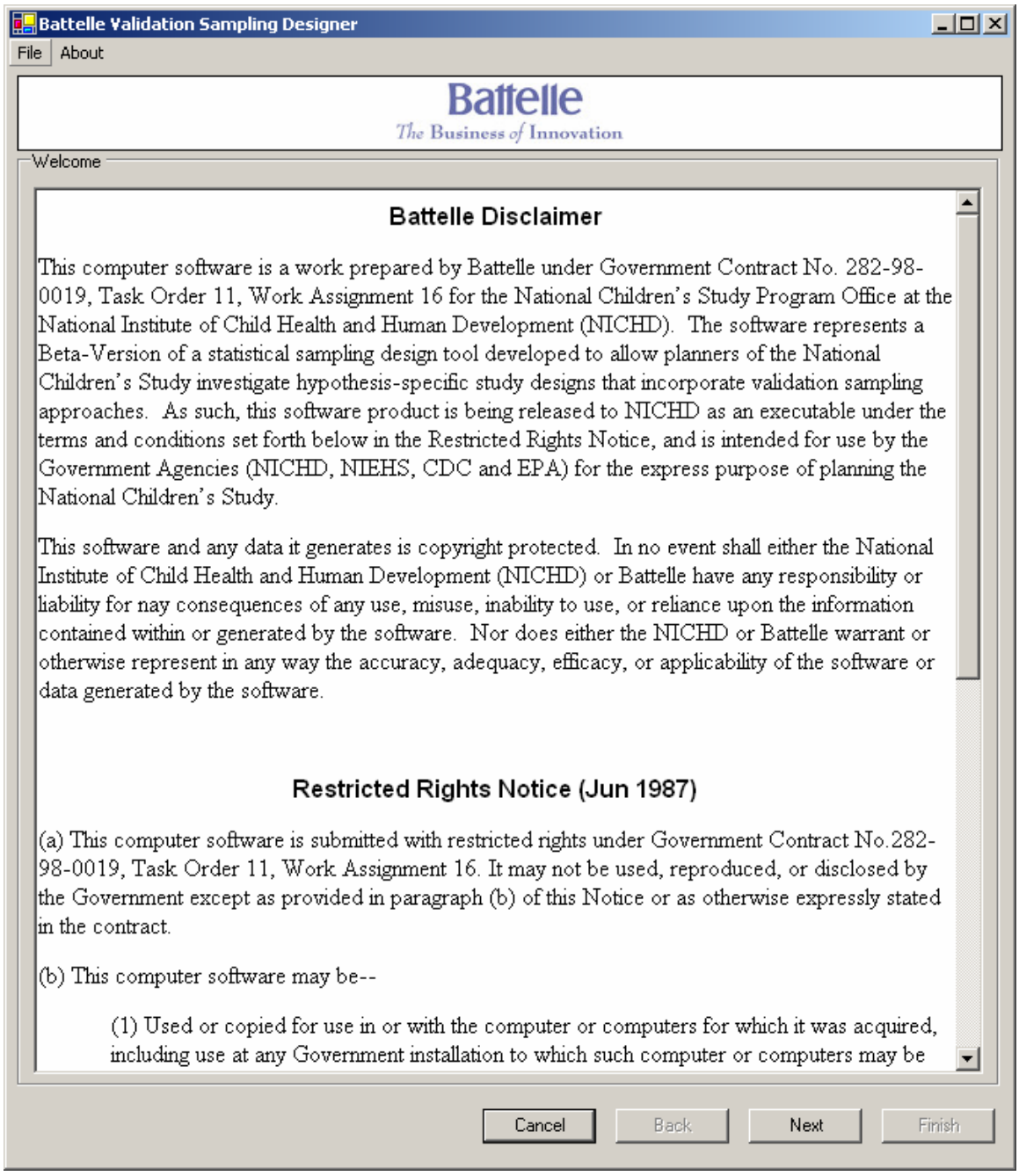


Figure 1 The Welcome Screen Provides the Battelle Disclaimer and the Restricted Rights Notice.

5.2 High Level Design Summary Window

After exiting the Welcome window, the next window allows the user to define the high-level characteristics about the design scenario that is being optimized.

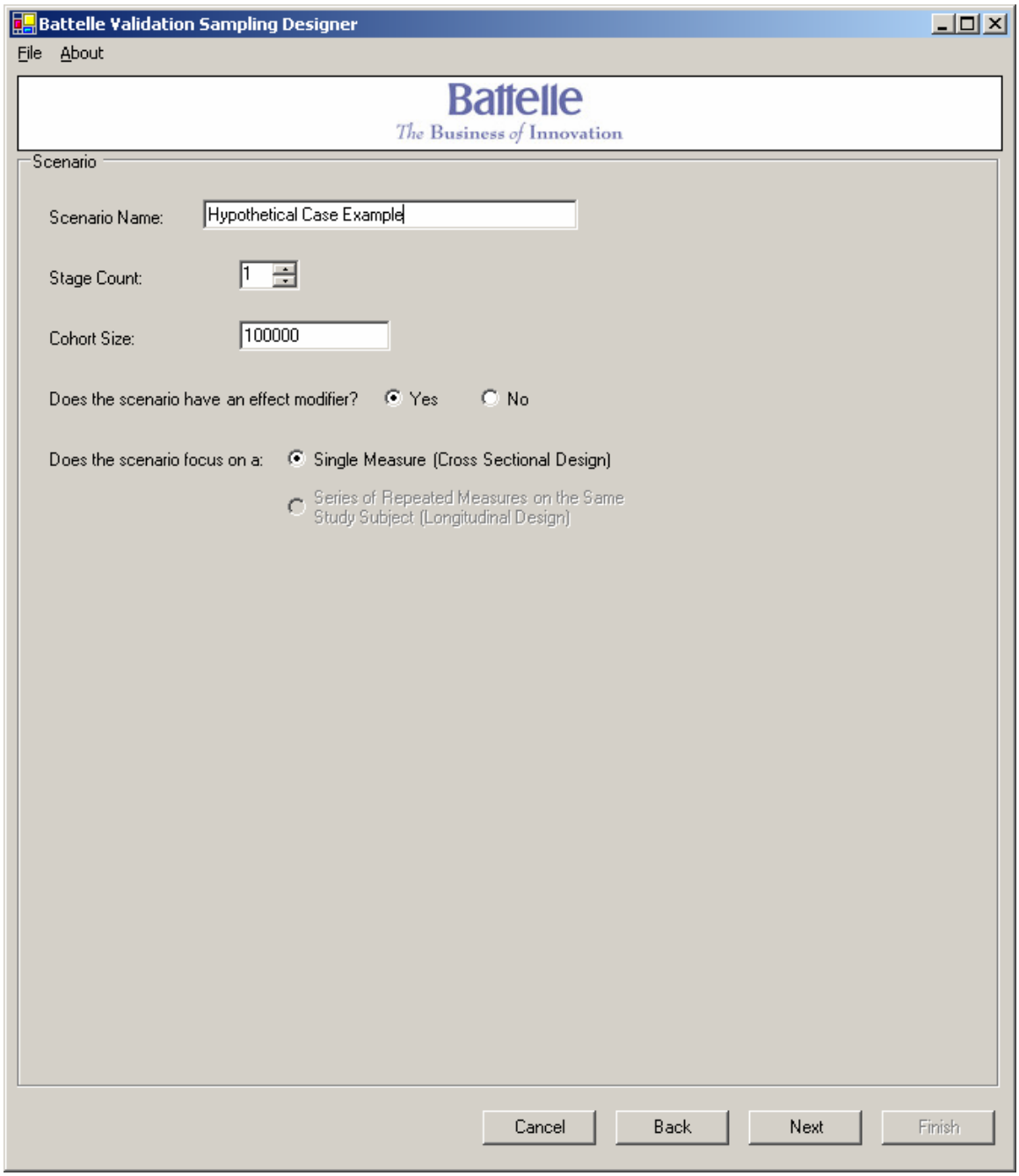


Figure 2 The High-Level Summary of Design Window

This window allows the user to provide the following input:

- ***Scenario Name:*** Type in a text description of the hypothesis being investigated
- ***Stage Count:*** This is an integer number ranging from 0 to 3 that represents the number of stages of surrogate measures that will be included in the design scenario
- ***Cohort Size:*** This is the number of study subjects that are available for sampling in Stage 0 of the design (the default value for the cohort size is set at 100,000 – corresponding to the number of anticipated live-birth outcomes for the National Children's Study). Stage 0 of the design corresponds to the fraction of study subjects that will be sampled for the health outcome. Note that the actual number of study subjects that will be included in Stage 0 may be a fraction of the total cohort size.
- ***Does the scenario have an effect modifier:*** This allows the user to determine whether or not to include an effect modifier in the design specifications. If the user selects to include an effect modifier, the graphical user interface will require the following input in subsequent windows:
    - The conditional distribution for the effect modifier as a function of the exposure variable (note that these variables could be assumed to be independent – in which case the user would be expected to provide information about the marginal distribution of the effect modifier).
    - The costs associated with sampling and laboratory analysis for the effect modifier.
    - The conditional distribution of Y|X,E – the distribution of the health outcome (Y)as a function of both the exposure variable (X) and the effect modifier (E).
- ***Does the scenario focus on a Single Measure (Cross Sectional Design) or a Series of Repeated Measures on the Same Study Subject (Longitudinal Design):*** This will allow the user (in subsequent versions of the software) to choose between a cross sectional and a repeated measures design. Currently, the software only allows for cross sectional designs.

In addition to these functions, at the top of this window (and all subsequent windows) is a drag-down File menu, which will allow the user to open a previously saved scenario, save the current scenario, or exit the software.

At the bottom of this window are three functional buttons: 'Next' which navigates the user to the next window in the structured interview; 'Back' which navigates the user to the previous window in the structured interview; and 'Cancel' which exits the user from the software. There is a fourth ('Finish') button at the bottom of the Window that will only become functional after the user has completed the structured interview process.

5.3 Exposure Variable Input Window

The Exposure Variable Input Window is designed to allow the user to provide input on the marginal distribution and costs associated with the main exposure variable of interest in the design, as seen in Figure 3.

The exposure variable is treated specially within the structured interview. It is the only variable whose distribution is input as a marginal distribution. The other variables in the structured interview (health outcome, effect modifier, and surrogate measures) are all defined as conditional distributions (as the function of the exposure variable, and perhaps other variables).

The user will be asked to provide a name for the exposure variable. This variable name will be used throughout the rest of the structured interview, and it is suggested to keep the name relatively short (e.g. X) - as the name will be used in several tables (and long variable names will force the user to scroll across the input screens).

The user will also be provided with three choices for the marginal distribution of the exposure variable:

- Normal: User provides the mean and standard deviation associated with the marginal distribution of the exposure variable.
- Binomial: User provides the prevalence associated with the marginal distribution of the exposure variable.
- Lognormal: User provides the geometric mean and geometric standard deviation associated with the marginal distribution of the exposure variable.

The user will also be asked whether (or not) the exposure variable is measurable.

If the exposure variable is measurable, it is automatically assumed that exposure will be measured in the last stage of sampling (and represents the most expensive and/or burdensome measure). When exposure is measurable, the user will be asked to provide the costs associated with sampling and analysis of exposure. The costs associated with analysis of the exposure measure are assumed to always occur in the last stage of sampling associated with the exposure measure. However, the costs associated with sampling for the exposure measure are allowed to be associated with any stage of sampling (because these costs may be necessary at previous stages if the user decides to explore either covariate or outcome dependent sampling designs). For the costs associated with sampling for the exposure variable, the user must select a stage in which those costs occur (even when the costs are zero). To force the user to provide this input, the default value for stage of sampling is listed as unknown, and must be changed by the user to proceed with the structured interview.

If the user selects that the exposure variable is not measurable, a latent variable model is assumed. This can be used when either the exposure variable is truly not measurable, or when the exposure variable is extremely expensive or burdensome and the study planner

would like to explore a design in which no study subjects undergo true exposure assessment. In the event that a latent variable model is assumed, the user must select either (1) two stages of surrogate measures, in which at least one of the surrogate measures follows the 'Classic Measurement Error Model' in its relationship with the exposure variable, or (2) three or more stages of surrogate measures must be utilized within the design. The 'Classic Measurement Error Model' is explained further in Section 5.6.

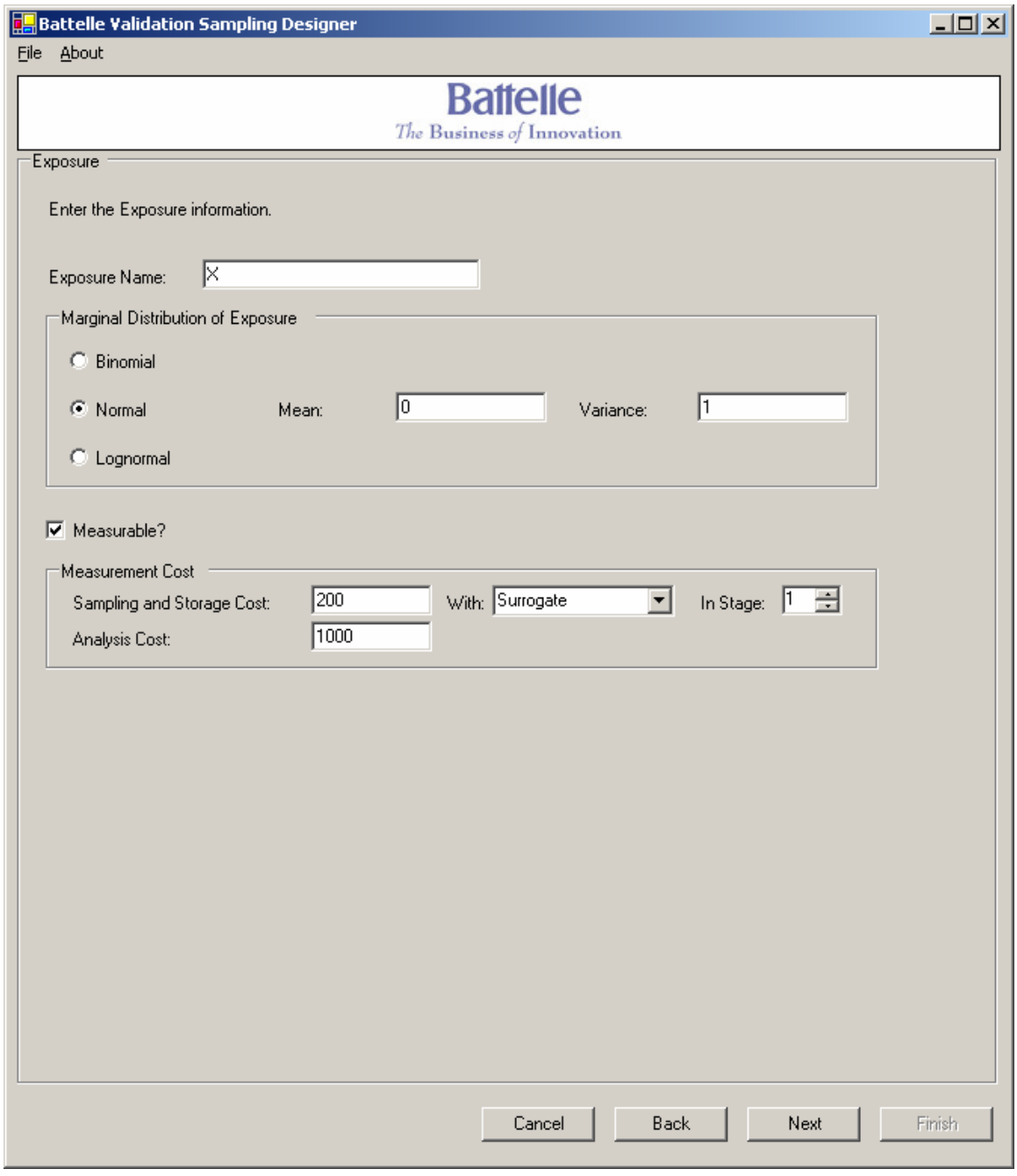


Figure 3 The Exposure Variable Input Window

At the bottom of this window are three buttons: 'Next' which navigates the user to the next window in the structured interview, 'Back' which navigates the user to the previous window in the structured interview, and 'Cancel' which exits the user from the software. There is a fourth ('Finish') button at the bottom of the Window that will only become functional after the user has completed the structured interview process.

5.4 Effect Modifier Input Window

The Effect Modifier Input Window, as seen in Figure 4, will only be presented to the user during the structured interview if the user had selected to include an effect modifier in the design as part of the input in the High Level Summary of Design Window (see Section 5.2).

The user will be asked to provide a name for the effect modifier variable. This variable name will be used throughout the rest of the structured interview, and it is suggested to keep the name relatively short (e.g. E) - as the name will be used in several tables (and long variable names will force the user to scroll across the input screens).

The user will then be asked whether there is any association between the effect modifier and the exposure variable, using a radio button. Subsections 5.4.1 and 5.4.2 provide details on inputting the distribution of the effect modifier, based on whether there is an association with the exposure variable.

Finally, the user will be asked to provide the costs associated with sampling and analysis of the effect modifier. The costs associated with analysis of the effect modifier are assumed to always occur in a stage of sampling immediately following the health outcome. However, the costs associated with sampling for the effect modifier is allowed to be associated with either the effect modifier stage, or the previous health outcome stage (because these costs may be necessary at previous stages if the user decides to explore an outcome dependent sampling design for the effect modifier). For the costs associated with sampling for the effect modifier, the user must select a stage in which those costs occur (even when the costs are zero). To force the user to provide this input, the default value for stage of sampling is listed as unknown, and must be changed by the user to proceed with the structured interview.

At the bottom of this window are three buttons: 'Next' which navigates the user to the next window in the structured interview, 'Back' which navigates the user to the previous window in the structured interview, and 'Cancel' which exits the user from the software. There is a fourth ('Finish') button at the bottom of the Window that will only become functional after the user has completed the structured interview process.

5.4.1 Inputting the Distribution of the Effect Modifier, Assuming No Association with the Exposure Variable

Assuming that there is no association between the effect modifier and exposure, the user will also be provided with three choices for the marginal distribution of the effect

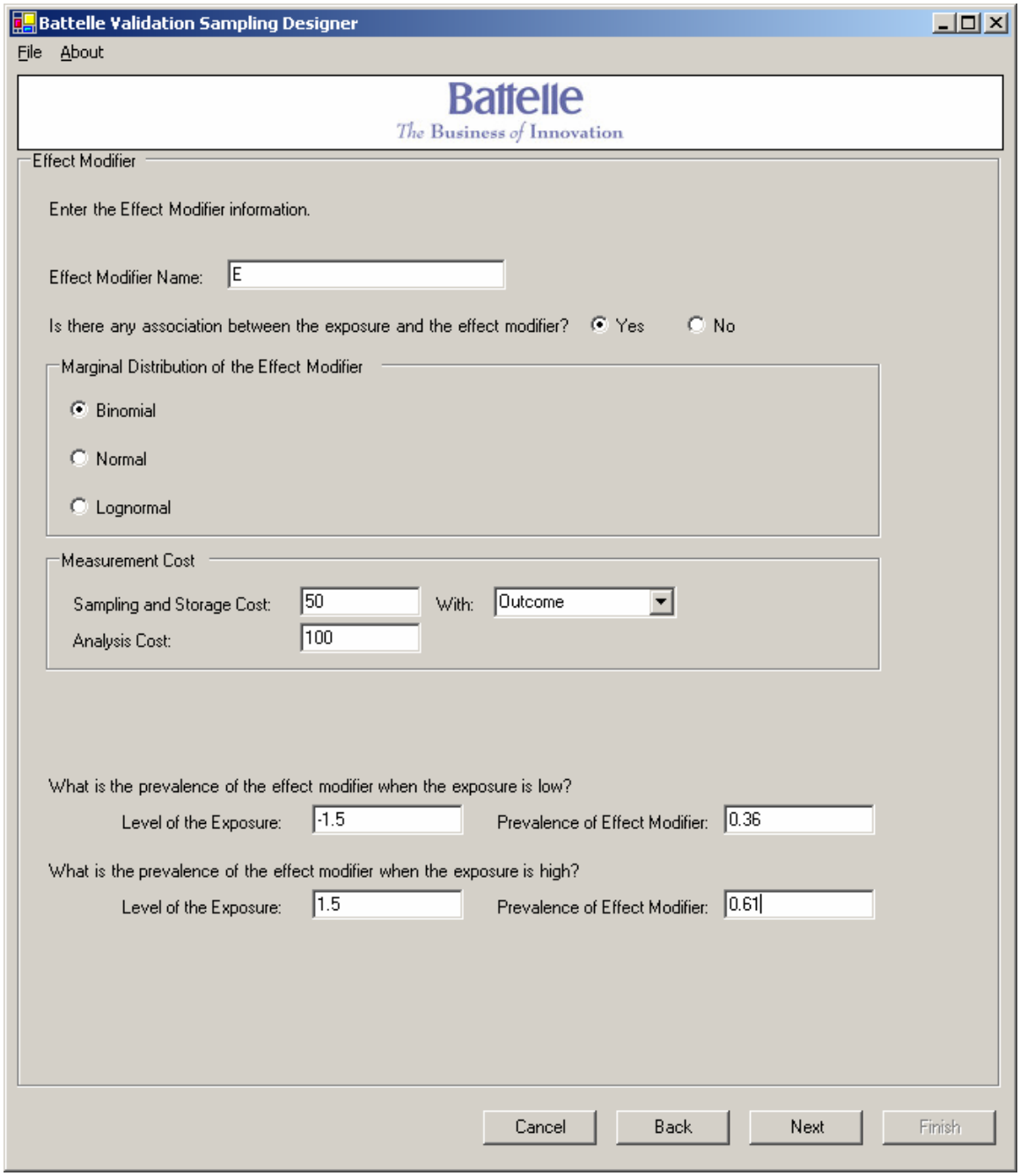


Figure 4 The Effect Modifier Input Window

modifier variable (similar to the choices provided earlier for the exposure variable):

- Normal: User provides the mean and standard deviation associated with the marginal distribution of the effect modifier.
- Binomial: User provides the prevalence associated with the marginal distribution of the effect modifier.
- Lognormal: User provides the geometric mean and geometric standard deviation associated with the marginal distribution of the effect modifier.

5.4.2 Inputting the Distribution of the Effect Modifier, Assuming an Association with the Exposure Variable

If there is an association between the effect modifier (E) and the exposure variable (X), the structured interview allows the user to provide statistical information that defines the conditional distribution of the effect modifier as a function of the exposure variable (E|X). The manner in which this information is gathered is based on the whether the previously entered marginal distribution of X is continuous or binary, and on whether E|X is continuous or binary. The following subsections provide details on how the conditional distribution of the effect modifier is ascertained based on input from the user:

5.4.2.1 Continuous Exposure / Continuous Effect Modifier

If both the exposure variable (X) and the effect modifier given the exposure variable (E|X) follow a continuous distribution (either Normal or Lognormal), then E|X is developed by inputting the marginal distribution of E and a measure of association (correlation coefficient) between E and X:

- X~Normal and E|X~Normal: User inputs the marginal distribution of E (mean and standard deviation), and the correlation coefficient between E and X ($\rho_{E,X}$).

- X~Normal and E|X~Lognormal: User inputs the marginal distribution of E (geometric mean and geometric standard deviation), and the correlation coefficient between ln(E) and X ($\rho_{\ln(E),X}$).

- X~Lognormal and E|X~Normal: User inputs the marginal distribution of E (mean and standard deviation), and the correlation coefficient between E and ln(X) ($\rho_{E,\ln(X)}$).

- X~Lognormal and E|X~Lognormal: User inputs the marginal distribution of E (geometric mean and geometric standard deviation), and the correlation coefficient between ln(E) and ln(X) ($\rho_{\ln(E),\ln(X)}$).

5.4.2.2 Continuous Exposure / Binary Effect Modifier

If the exposure variable (X) follows a continuous distribution (either Normal or Lognormal), and the effect modifier given the exposure variable (E|X) follows a binomial distribution, then E|X is developed by providing input to the following two questions:

- What is the prevalence of the effect modifier (E) when exposure (X) is low?
  - User provides a target low level for the exposure variable (X)
  - User provides the prevalence of the effect modifier, when X is equal to the above target level.

- What is the prevalence of the effect modifier (E) when exposure (X) is high?
  - User provides a target high level for the exposure variable (X)
  - User provides the prevalence of the effect modifier (E), when X is equal to the above target level.

The target levels (low and high) for exposure (X) that are input by the user for the above two questions should ideally be values that are away from the median of the marginal distribution of X (i.e. the low target value should be below the 25^{th} percentile of the marginal distribution of X, and the high target value should be above the 75^{th} percentile of the marginal distribution of X). If the user provides target values that are too close to the median of the marginal distribution of X, the structured interview will prompt the user for another value (if possible).

5.4.2.3 Binary Exposure / Continuous Effect Modifier

If the exposure variable (X) follows a binomial distribution, and the effect modifier given the exposure variable (E|X) follows a continuous distribution (either Normal or Lognormal), then E|X is developed by providing input to any two of the following three questions:

- What is the marginal distribution of the effect modifier (E)?
  - If E follows a Normal distribution, then user provides the mean and standard deviation.
  - If E follows a Lognormal distribution, then user provides the geometric mean and geometric standard deviation.

- What is the mean value of the effect modifier (E) when exposure (X) is low?
  - When X follows a binomial distribution, it is automatically assumed that exposure is low when X=0.
  - User provides the mean value of the effect modifier, when X=0.

- What is the mean value of the effect modifier (E) when exposure (X) is high?
  - When X follows a binomial distribution, it is automatically assumed that exposure is high when X=1.

- User provides the mean value of the effect modifier, when X=1.

The structured interview will only allow the user to provide input to two of the above three questions.

5.4.2.4 Binary Exposure / Binary Effect Modifier

If both the exposure variable (X) and the effect modifier given the exposure variable (E|X) follow a binomial distribution (either Normal or Lognormal), then E|X is defined by the user by inputting both (1) the prevalence of the marginal distribution of the effect modifier (E), and (2) the odds ratio between the effect modifier (E) and the exposure variable (X) (ψ_{EX}).

.

5.5 Health Outcome Input Window

The user will be asked to provide a name for the health outcome variable, which is assumed to be measured in the first stage of sampling. This variable name will be used throughout the rest of the structured interview, and it is suggested to keep the name relatively short (e.g. Y) - as the name will be used in several tables (and long variable names will force the user to scroll across the input screens).

The user will then be asked for information that defines the distribution of the health outcome (Y) as a function of exposure (X) (and possibly the effect modifier, E). Subsection 5.5.1 provides details on inputting the distribution of Y|X (no effect modifier), and Subsection 5.5.1 provides details on inputting the distribution of Y|X,E.

Finally, the user will be asked to provide the costs associated with sampling and analysis of the health outcome variable. Unlike the other variables that are included in the structured interview, these costs are always assumed to be measured in the first stage of sampling (thus, the sampling costs cannot be assigned to a previous stage).

At the bottom of this window are three buttons: 'Next' which navigates the user to the next window in the structured interview, 'Back' which navigates the user to the previous window in the structured interview, and 'Cancel' which exits the user from the software. There is a fourth ('Finish') button at the bottom of the Window that will only become functional after the user has completed the structured interview process.

An example of the Health Outcome Input Window is provided in Figure 5, for a design scenario with a continous exposure variable and a binary health outcome measure (with no effect modifier).

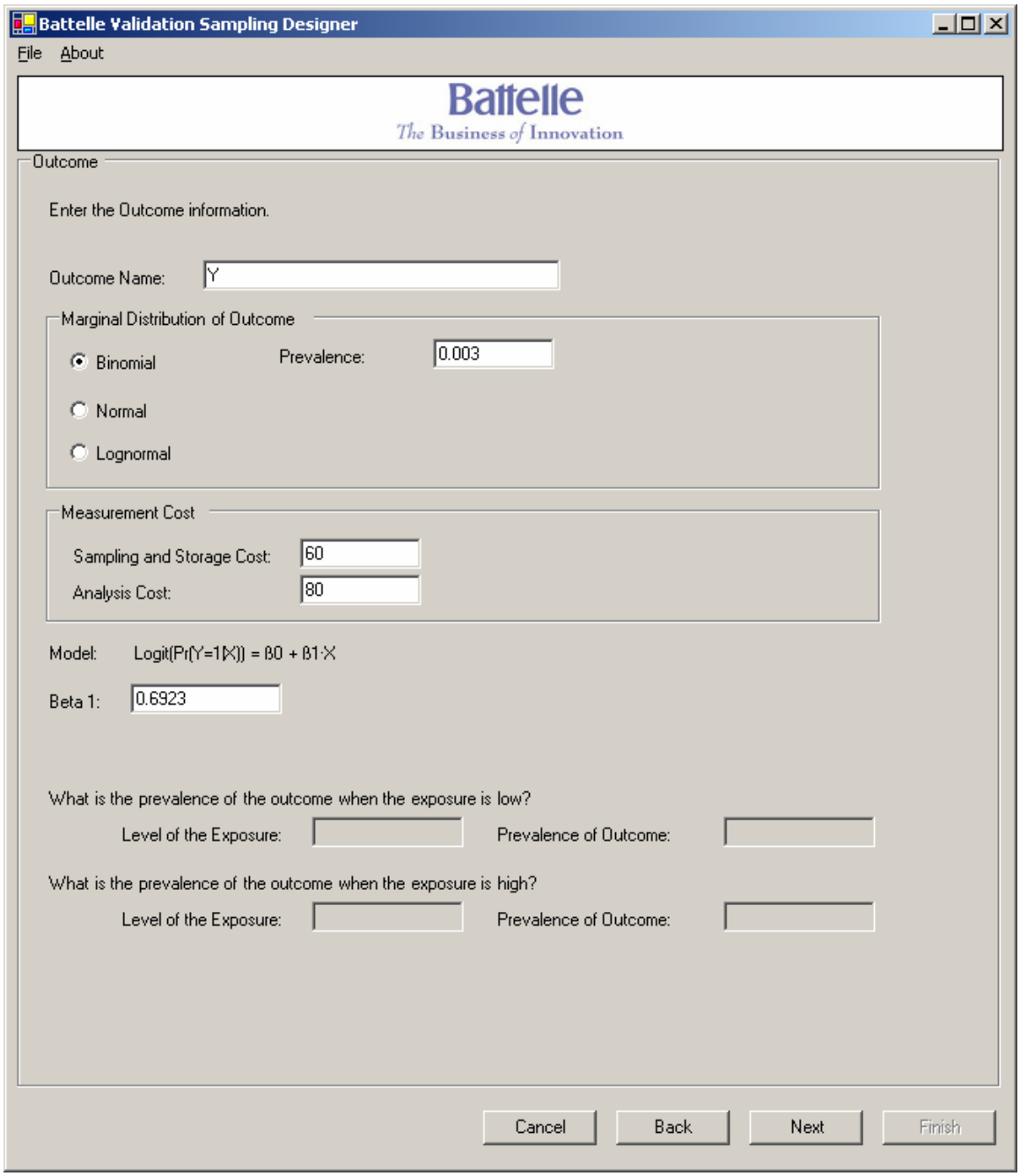


Figure 5 The Health Outcome Input Window (No Effect Modifier)

5.5.1 Inputting the Conditional Distribution of the Health Outcome as a Function of Exposure (Y|X), Assuming No Effect Modifier.

If there is no effect modifier, the structured interview allows the user to provide statistical information that defines the conditional distribution of the health outcome as a function of the exposure variable (Y|X). The manner in which this information is gathered is based on the whether the previously entered marginal distribution of X is continuous or

binary, and on whether Y|X is continuous or binary. The following subsections provide details on how the conditional distribution of the health outcome is ascertained based on input from the user:

5.5.1.1 Continuous Health Outcome

If the health outcome given the exposure variable (Y|X) follows a continuous distribution (either Normal or Lognormal), then Y|X is developed by inputting the marginal distribution of Y and a measure of association between Y and X:

- Marginal Distribution of Y:
  - If Y|X is assumed to follow a Normal distribution, the user provides the mean and standard deviation for the marginal distribution of Y.
  - If Y|X is assumed to follow a Lognormal distribution, the user provides the geometric mean and geometric standard deviation for the marginal distribution of Y.

- Measure of Association between Y and X: The relationship between Y and X can be defined in one of the following three ways (these are statistically equivalent pieces of information):
  - Provide a value of β_1 from the regression relationship of Y|X
    - $Y = \beta_0 + \beta_1 \cdot X + \text{Error}$ – If Y|X follows a Normal distribution and X follows a Normal or Binomial distribution
    - $Y = \beta_0 + \beta_1 \cdot \ln(X) + \text{Error}$ – If X follows a Lognormal distribution and Y|X follows a Normal distribution
    - $\ln(Y) = \beta_0 + \beta_1 \cdot X + \text{Error}$ – If X follows a Normal or Binomial distribution and Y|X follows a Lognormal distribution
    - $\ln(Y) = \beta_0 + \beta_1 \cdot \ln(X) + \text{Error}$ – If both X and Y|X follow a Lognormal distribution
    - Note that the analyzer will not engage if the marginal variance of Y is less than the variability assumed in the above regression relationships (based on the variance of X and the β_1 coefficient entered by the user).

  - Provide a value of the variance of Y|X
    - This is assumed to be less than the variance defined earlier within the marginal distribution of Y. If the user provides a value greater than or equal to the marginal variance of Y, the software will prompt the user for a smaller value.
    - Since the user previously provided the marginal variance of Y, knowing the variance of Y|X will allow the software to determine the fraction of variability in Y explained by X (R^2).

  - Provide the mean value of the health outcome (Y) when exposure (X) is low (or high)?

- User provides a target high level for the exposure variable (X). This value should be away from the median of the distribution of X (lower than the 25th percentile or higher than the 75th percentile).

5.5.1.2 Binary Health Outcome

If the health outcome given the exposure variable (Y|X) follows a binomial distribution, then Y|X is developed by providing input in one of the following two manners:

- Provide the prevalence of the marginal distribution of Y and the value of β_1 from the following logistic regression model:
  - $\text{Logit}(\pi) = \beta_0 + \beta_1 \cdot X$, where $\pi = \Pr(Y=1)$ and X follows a Normal or Binomial distribution
  - $\text{Logit}(\pi) = \beta_0 + \beta_1 \cdot \ln(X)$, where $\pi = \Pr(Y=1)$ and X follows a Lognormal distribution

- Provide answers to the following two questions:
  - What is the prevalence of the effect modifier (E) when exposure (X) is low?
    - User provides a target low level for the exposure variable (X). When X follows a binary distribution, exposure is assumed to be low when X=0.
    - User provides the prevalence of the effect modifier, when X is equal to the above target level.

  - What is the prevalence of the effect modifier (E) when exposure (X) is high?
    - User provides a target high level for the exposure variable (X). When X follows a binary distribution, exposure is assumed to be high when X=1.
    - User provides the prevalence of the effect modifier (E), when X is equal to the above target level.

  The target levels (low and high) for continuous exposure (X) that are input by the user for the above two questions should ideally be values that are away from the median of the marginal distribution of X (i.e. the low target value should be below the 25th percentile of the marginal distribution of X, and the high target value should be above the 75th percentile of the marginal distribution of X). If the user provides target values that are too close to the median of the marginal distribution of X, the structured interview will prompt the user for another value (if possible).

5.5.2 Inputting the Conditional Distribution of the Health Outcome as a Function of Exposure and Effect Modifier (Y|X,E)

If there is an effect modifier (based on previous input as described in Sections 5.2 and 5.4), the structured interview allows the user to provide statistical information that defines the conditional distribution of the health outcome as a function of the exposure variable and the effect modifier (Y|X,E). The manner in which this information is gathered is based on the whether the previously entered marginal distribution of X is continuous or binary, on whether the conditional distribution of E|X is continuous or binary, and on whether Y|X,E is continuous or binary. The following subsections provide details on how the conditional distribution of the health outcome is ascertained based on input from the user:

5.5.2.1 Continuous Health Outcome

If the health outcome given the exposure variable and effect modifier (Y|X,E) follows a continuous distribution (either Normal or Lognormal), then Y|X is developed by inputting the marginal distribution of Y, a measure of R^2 (the fraction of variability in the marginal distribution of Y explained by X and E), and the β_1, β_2 and β_3 coefficients from one of the following linear regression models:

- $Y= \beta_0 + \beta_1 \cdot E + \beta_2 \cdot X + \beta_3 \cdot E \cdot X + Error$ – If Y|X follows a Normal distribution and both X and E follow either a Normal or Binomial distribution;
- $Y= \beta_0 + \beta_1 \cdot \ln(E) + \beta_2 \cdot X + \beta_3 \cdot \ln(E) \cdot X + Error$ – If Y|X follows a Normal distribution, X follows either a Normal or Binomial distribution, and E follows a Lognormal distribution;
- $Y= \beta_0 + \beta_1 \cdot E + \beta_2 \cdot \ln(X) + \beta_3 \cdot E \cdot \ln(X) + Error$ – If Y|X follows a Normal distribution, X follows a Lognormal distribution, and E follows either a Normal or Binomial distribution;
- $Y= \beta_0 + \beta_1 \cdot \ln(E) + \beta_2 \cdot \ln(X) + \beta_3 \cdot \ln(E) \cdot \ln(X) + Error$ – If Y|X follows a Normal distribution and both X and E follow a Lognormal distribution;
- $\ln(Y)= \beta_0 + \beta_1 \cdot E + \beta_2 \cdot X + \beta_3 \cdot E \cdot X + Error$ – If Y|X follows a Lognormal distribution and both X and E follow either a Normal or Binomial distribution;
- $\ln(Y)= \beta_0 + \beta_1 \cdot \ln(E) + \beta_2 \cdot X + \beta_3 \cdot \ln(E) \cdot X + Error$ – If Y|X follows a Lognormal distribution, X follows either a Normal or Binomial distribution, and E follows a Lognormal distribution;
- $\ln(Y)= \beta_0 + \beta_1 \cdot E + \beta_2 \cdot \ln(X) + \beta_3 \cdot E \cdot \ln(X) + Error$ – If Y|X follows a Lognormal distribution, X follows a Lognormal distribution, and E follows either a Normal or Binomial distribution;
- $\ln(Y)= \beta_0 + \beta_1 \cdot \ln(E) + \beta_2 \cdot \ln(X) + \beta_3 \cdot \ln(E) \cdot \ln(X) + Error$ – If Y|X follows a Lognormal distribution and both X and E follow a Lognormal distribution;

Note that the analyzer will not engage if the marginal variance of Y is less than the variability assumed in the above regression relationships (based on the joint variance of X and E, and the β coefficients entered by the user).

5.5.2.2 Binary Health Outcome

If the health outcome given the exposure variable and effect modifier (Y|X,E) follows a binomial distribution, then Y|X is developed by inputting the marginal distribution of Y (the prevalence), and the β_1, β_2 and β_3 coefficients from one of the following logistic regression models:

- $\text{logit}(\pi) = \beta_0 + \beta_1 \cdot E + \beta_2 \cdot X + \beta_3 \cdot E \cdot X$ -- where $\pi = \Pr(Y=1)$, and both X and E follow either a Normal or Binomial distribution;
- $\text{logit}(\pi) = \beta_0 + \beta_1 \cdot \ln(E) + \beta_2 \cdot X + \beta_3 \cdot \ln(E) \cdot X$ -- where $\pi = \Pr(Y=1)$, X follows either a Normal or Binomial distribution, and E follows a Lognormal distribution;
- $\text{logit}(\pi) = \beta_0 + \beta_1 \cdot E + \beta_2 \cdot \ln(X) + \beta_3 \cdot E \cdot \ln(X)$ -- where $\pi = \Pr(Y=1)$, X follows a Lognormal distribution, and E follows either a Normal or Binomial distribution;
- $\text{logit}(\pi) = \beta_0 + \beta_1 \cdot \ln(E) + \beta_2 \cdot \ln(X) + \beta_3 \cdot \ln(E) \cdot \ln(X)$ -- where $\pi = \Pr(Y=1)$, and both X and E follow a Lognormal distribution.

Figure 6 provides an example of the Health Outcome Input Window for a logistic regression model which involves a lognormal exposure variable and a binary effect modifier.

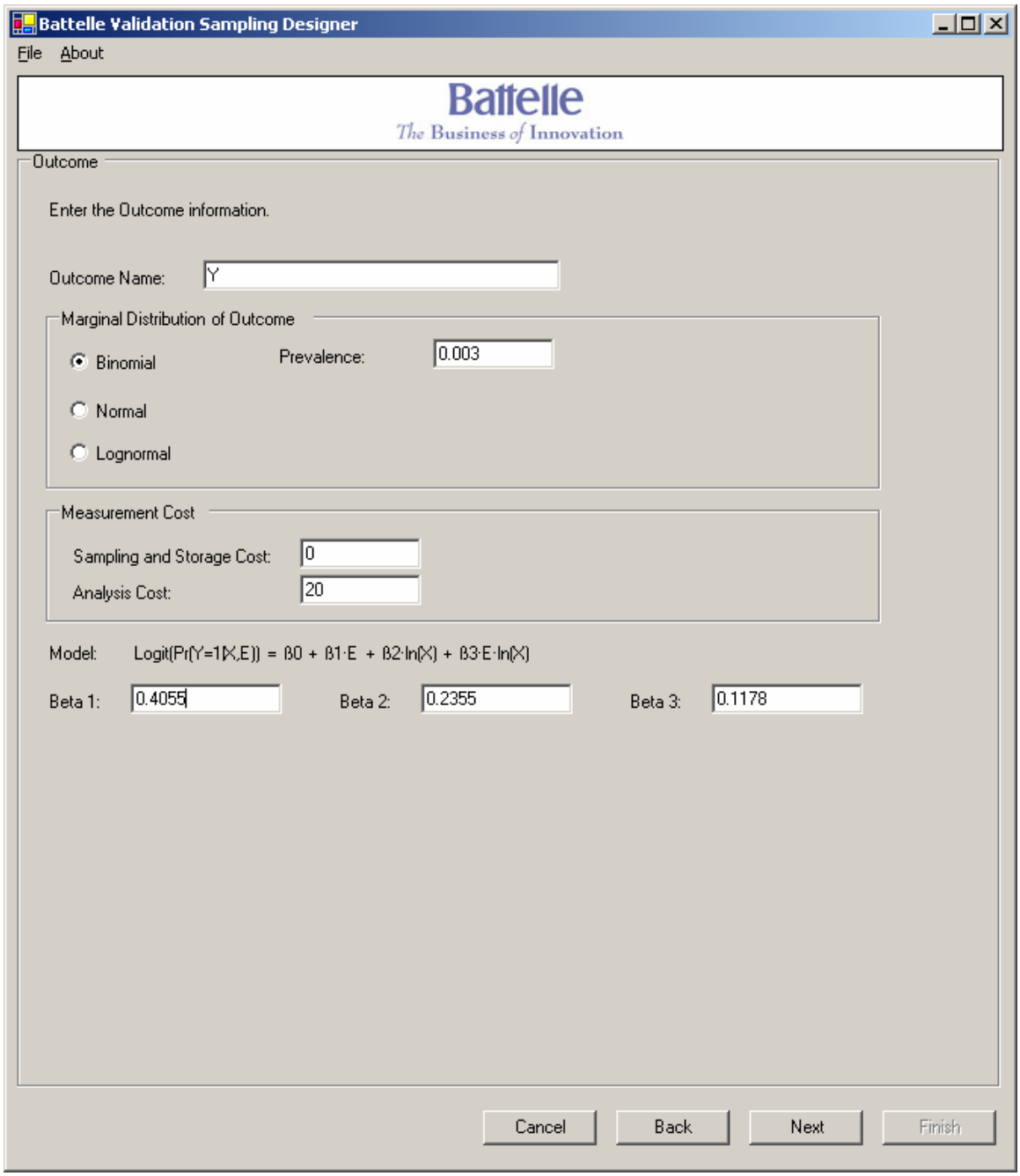


Figure 6 The Health Outcome Input Window (with Effect Modifier)

5.6 Inputting Surrogate Measures Windows

The next window in the structured interview provides a gateway for the user to provide input on one or more candidate surrogate measures of exposure for the validation sampling design. As seen in Figure 7, this gateway window provides a listing of all

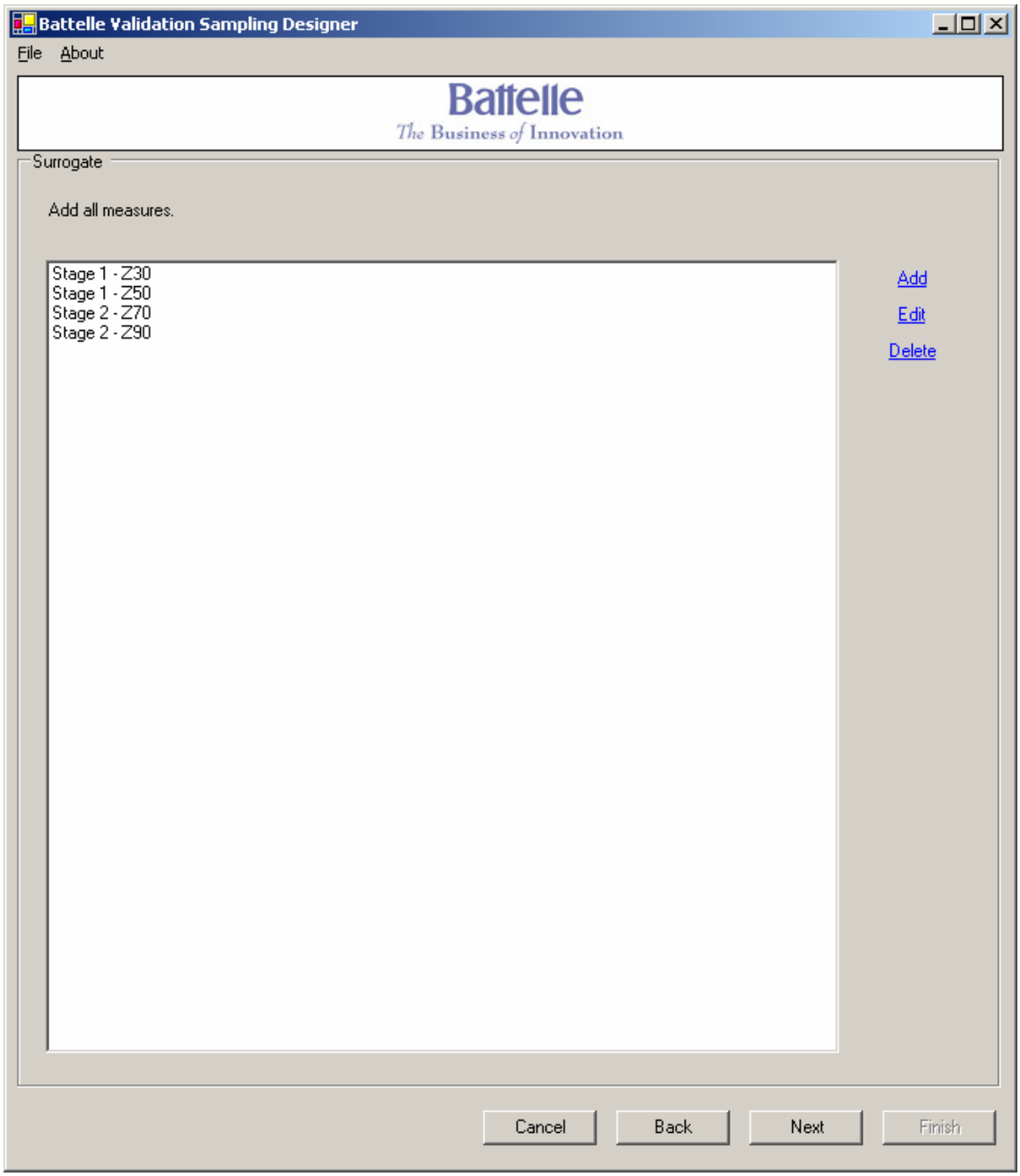


Figure 7 The Gateway Window for Inputting Surrogate Measures

previously entered candidate surrogate exposure measures (by stage of sampling) on the left-hand side, and buttons that allow the user to either Add a new measure, or Edit or Delete an existing surrogate measure.

The user can choose to enter multiple candidate surrogate measures of exposure (with different properties in terms of cost and relationship with the exposure variable) within the same stage of the design. In this case, the software will automatically explore every combination of variables that can be used within the design. For example, if the user is exploring designs with two stages of surrogate sampling and two candidate surrogate exposure measures are entered for each stage (as seen in Figure 7), then the software will develop optimal validation sampling designs for all four possible ways of combining the candidate measures in a two stage design (Z30/Z70, Z30/Z90, Z50/Z70, and Z50/Z90).

Subsection 5.6.1 provides details on cost and distributional input that is solicited once the user selects to Add a new surrogate measure (or Edit an existing surrogate measure).

It should be noted that after completion of this part of the structured interview, the user should make sure that the following checks have been completed (otherwise, the analyzer may not engage, or it may yield nonsensical results):

1. There is at least one surrogate measure entered for each stage of surrogate sampling that was included in the design input from the first "High Level Summary of Design" Window.
2. If the user defined the exposure variable as latent (not measurable) – then the user must select either (1) two stages of surrogate measures, in which at least one of the surrogate measures follows the 'Classic Measurement Error Model' in its relationship with the exposure variable, or (2) three or more stages of surrogate measures must be utilized within the design.

At the bottom of this gateway window are three buttons: 'Next' which navigates the user to the next window in the structured interview; 'Back' which navigates the user to the previous window in the structured interview; and 'Cancel' which exits the user from the software. There is a fourth ('Finish') button at the bottom of the Window that will only become functional after the user has completed the structured interview process.

5.6.1 Adding/Editing Surrogate Measures Window

Once the user selects to Add a new surrogate measure of exposure (or Edit an existing surrogate measure of exposure), a new Window will appear that allows the user to provide the appropriate input on the costs associated with the surrogate measure of exposure and the conditional distribution of the surrogate measure of exposure as a function of the exposure variable (as seen in Figure 8).

The user will be asked to provide a name for the surrogate measure of exposure. This variable name will be used throughout the rest of the structured interview, and it is suggested to keep the name relatively short (e.g. Z1) - as the name will be used in several

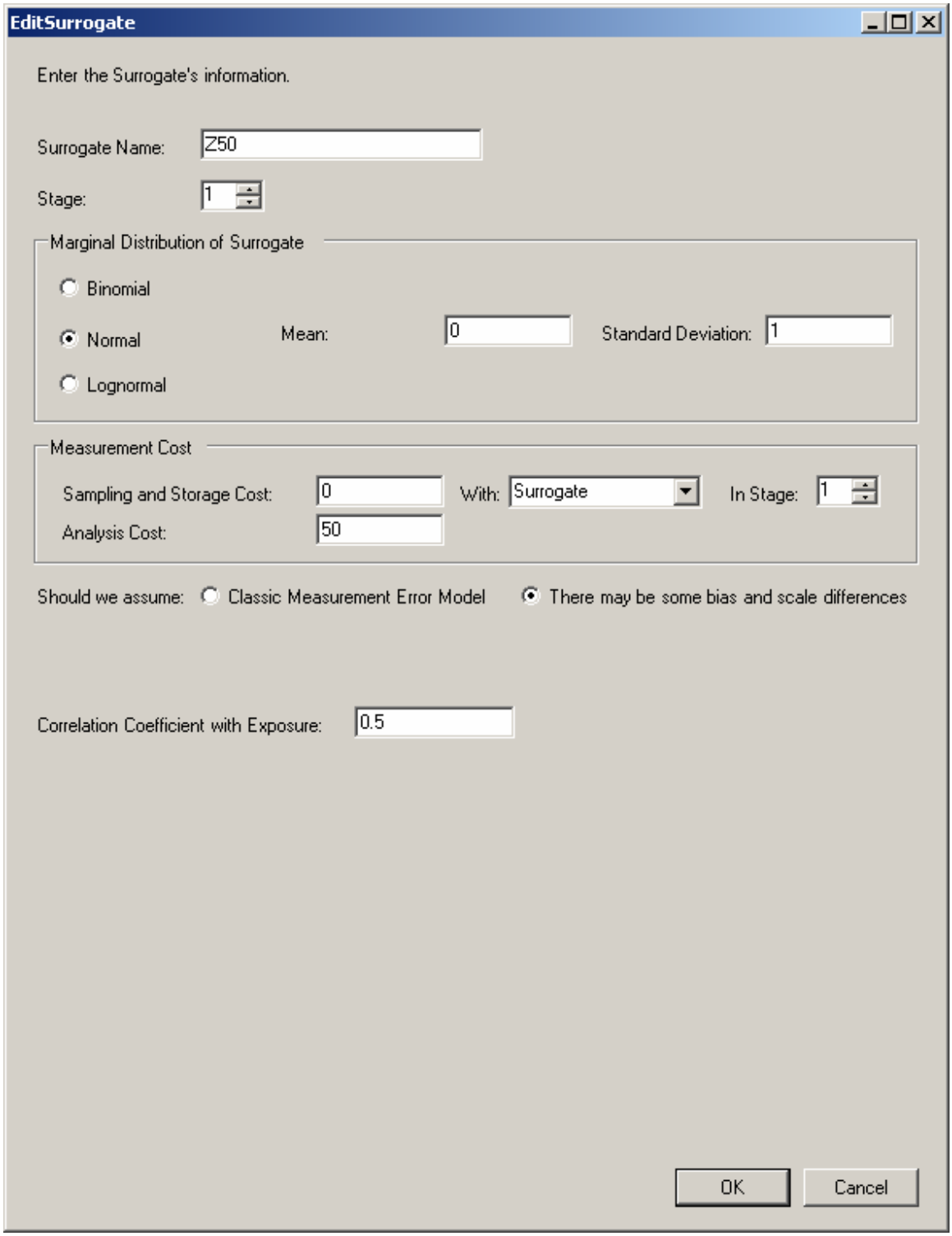


Figure 8 Example Adding/Editing Surrogate Measures Window

tables (and long variable names will force the user to scroll across the input screens). The user will also be asked to identify the stage of surrogate sampling associated with this particular measure in the design.

The user will then be asked to provide information that defined the conditional distribution of the surrogate measure of exposure as a function of the exposure variable, as described in Subsections 5.6.1.1 – 5.6.1.4. In these Subsections, we will refer to the surrogate measure of exposure as Z, and the exposure variable as X.Finally, the user will be asked to provide the costs associated with sampling and analysis of the surrogate measure of exposure. The costs associated with analysis of the surrogate measure of exposure are assumed to always occur in the stage of surrogate sampling identified within this Window. However, the costs associated with sampling for the surrogate measure of exposure is allowed to be associated with the stage of sampling identified, or any previous stage of sampling (because these costs may be necessary at previous stages if the user decides to explore a covariate or outcome dependent sampling design for the surrogate measure of exposure). For the costs associated with sampling of the surrogate measure of exposure, the user must select a stage in which those costs occur (even when the costs are zero). To force the user to provide this input, the default value for stage of sampling is listed as unknown, and must be changed by the user to proceed with the structured interview.

At the bottom of the Adding/Editing Surrogate Measures Window are two buttons: 'OK' which accepts the input from the current Window and navigates the user back to the Gateway Window for Inputting Surrogate Measures, and 'Cancel' which navigates the user back to the Gateway Window for Inputting Surrogate Measures without accepting the input from the current Window.

5.6.1.1 Defining Z|X When Both Z and X are Continuous

If both the exposure variable (X) and the surrogate measure of exposure given the exposure variable (Z|X) follow a continuous distribution (either Normal or Lognormal), then Z|X is developed in one of two ways:

1. Assume that there may be some bias and/or scale differences between the surrogate measure of exposure and the exposure variable (use the radio button to make this selection). This implies that the relationship between Z and X (assuming both follow a normal distribution) can be defined by the regression relationship $Z = \beta_0 + \beta_1 \cdot X + \text{Error}$, where $\beta_0 \neq 0$ and/or $\beta_0 \neq 1$. The user will then input the marginal distribution of Z and a measure of association (correlation coefficient) between Z and X:

    - X~Normal and E|X~Normal: User inputs the marginal distribution of E (mean and standard deviation), and the correlation coefficient between E and X ($\rho_{E,X}$).

    - X~Normal and E|X~Lognormal: User inputs the marginal distribution of E (geometric mean and geometric standard deviation), and the correlation coefficient between ln(E) and X ($\rho_{\ln(E),X}$).

- X~Lognormal and E|X~Normal: User inputs the marginal distribution of E (mean and standard deviation), and the correlation coefficient between E and ln(X) ($\rho_{E,ln(X)}$).

- X~Lognormal and E|X~Lognormal: User inputs the marginal distribution of E (geometric mean and geometric standard deviation), and the correlation coefficient between ln(E) and ln(X) ($\rho_{ln(E),ln(X)}$).

2. Assume the 'Classic Measurement Error Model' (use the radio button to make this selection). This implies that there are no bias or scale differences between the surrogate measure of exposure and the exposure variable, and that the surrogate measure is just an error-prone version of the exposure variable as defined by the regression relationship Z = X + Error (assuming both follow a normal distribution). The user will then input a measure of association (correlation coefficient) between Z and X:

    - X~Normal and E|X~Normal: User inputs the correlation coefficient between E and X ($\rho_{E,X}$).

    - X~Normal and E|X~Lognormal: User inputs the correlation coefficient between ln(E) and X ($\rho_{ln(E),X}$).

    - X~Lognormal and E|X~Normal: User inputs the correlation coefficient between E and ln(X) ($\rho_{E,ln(X)}$).

    - X~Lognormal and E|X~Lognormal: User inputs the correlation coefficient between ln(E) and ln(X) ($\rho_{ln(E),ln(X)}$).

Please note that the assumption of the 'Classic Measurement Error Model' as described in item (2) above is a strong assumption within the context of Validation Sampling Design. This assumption removes degrees of freedom from the design optimization – and suggests that the joint distribution between Z and X can be characterized by 3 parameters rather than 5 parameters. This assumption is also necessary for designs in which the user selects the exposure variable as latent (not measurable) and only 2 stages of surrogate exposure measures (in which case, one of the surrogate measures of exposure must follow the 'Classic Measurement Error Model').

5.6.1.2 Defining Z|X When Z is Binary and X is Continuous

If the exposure variable (X) follows a continuous distribution (either Normal or Lognormal), and the surrogate measure of exposure given the exposure variable (Z|X) follows a binomial distribution, then Z|X is developed by providing input to two of the following three questions:

- What is the marginal prevalence of the surrogate measure of exposure (Z)?

- What is the prevalence of the surrogate measure (Z) when exposure (X) is low?
  - User provides a target low level for the exposure variable (X)
  - User provides the prevalence of the surrogate measure (Z), when X is equal to the above target level.

- What is the prevalence of the surrogate measure (Z) when exposure (X) is high?
  - User provides a target high level for the exposure variable (X)
  - User provides the prevalence of the surrogate measure (Z), when X is equal to the above target level.

The target levels (low and high) for exposure (X) that are input by the user for the second two questions above should ideally be values that are away from the median of the marginal distribution of X (i.e. the low target value should be below the 25^{th} percentile of the marginal distribution of X, and the high target value should be above the 75^{th} percentile of the marginal distribution of X). If the user provides target values that are too close to the median of the marginal distribution of X, the structured interview will prompt the user for another value (if possible).

5.6.1.3 Defining Z|X When Z is Continuous and X is Binary

If the surrogate measure of exposure given the exposure variable (Z|X) follows a continuous distribution (either Normal or Lognormal) and X follows a binomial distribution, then Z|X is developed by inputting the marginal distribution of Z and a measure of association between Z and X:

- Marginal Distribution of Z:
  - If Z|X is assumed to follow a Normal distribution, the user provides the mean and standard deviation for the marginal distribution of Z.
  - If Z|X is assumed to follow a Lognormal distribution, the user provides the geometric mean and geometric standard deviation for the marginal distribution of Z.

- Measure of Association between Z and X: The relationship between Z and X can be defined in one of the following three ways (these are statistically equivalent pieces of information):
  - Provide a value of β_1 from the regression relationship of Z|X
    - $Z = \beta_0 + \beta_1 \cdot X + \text{Error}$ – If Z|X follows a Normal distribution
    - $\ln(Z) = \beta_0 + \beta_1 \cdot X + \text{Error}$ – If Z|X follows a Lognormal distribution
    - Note that the analyzer will not engage if the marginal variance of Y is less than the variability assumed in the above regression relationships (based on the variance of X and the β_1 coefficient entered by the user).

  - Provide a value of the variance of Z|X or the R^2 Coefficient

- If the user selects to provide the variance of Z|X, the inputted value must be less than the marginal variance of Z. If the user provides a value greater than or equal to the marginal variance of Z, the software will prompt the user for a smaller value.
- If the user selects to provide the R^2 coefficient, the value must be greater than zero and less than one.

- Provide the mean value of the health outcome (Z) when exposure (X) is low (or high)?

5.6.1.4 Defining Z|X When Both Z and X are Binary

If the both the surrogate measure of exposure given the exposure variable (Z|X) and the exposure variable (X) follow a binomial distribution, then the distribution of Z|X is developed by providing input in one of the following two manners:

- Provide the prevalence of the marginal distribution of Z, and the value of the odds ratio between the Z and X ($\psi_{Z,X}$)

- Provide answers to the following two questions:
  - What is the prevalence of the surrogate measure of exposure (Z) when exposure (X) is low (X=0)?
  - What is the prevalence of the surrogate measure of exposure (Z) when exposure (X) is high (X=1)?

5.7 The Linked Stages of Sampling Constraints Window

The next step in the structured interview allows the user to input constraints on the design so that any candidate variable in the design can be sampled for the same subset of study participants that are included in the previous stage of sampling. Figure 9 provides an example of Linked Stages of Sampling, in which there is a health outcome (Y), two stages of surrogate measures (Z50 and Z90), and an exposure variable (X). In this example design, a box is checked that states "Always sample Z50 with the previous stage." This constrains the design so that the surrogate exposure variable Z50 is measured on the identical subset of study participants that have the health outcome (Y) assessed.

If the user would like for the health outcome (Y) to be assessed on every study subject in the cohort, then the first line (Always sample Y with the previous stage) should be checked.

At the bottom of this window are three buttons: 'Next' which navigates the user to the next window in the structured interview, 'Back' which navigates the user to the previous

window in the structured interview, and 'Cancel' which exits the user from the software. There is a fourth ('Finish') button at the bottom of the Window that will only become functional after the user has completed the structured interview process.

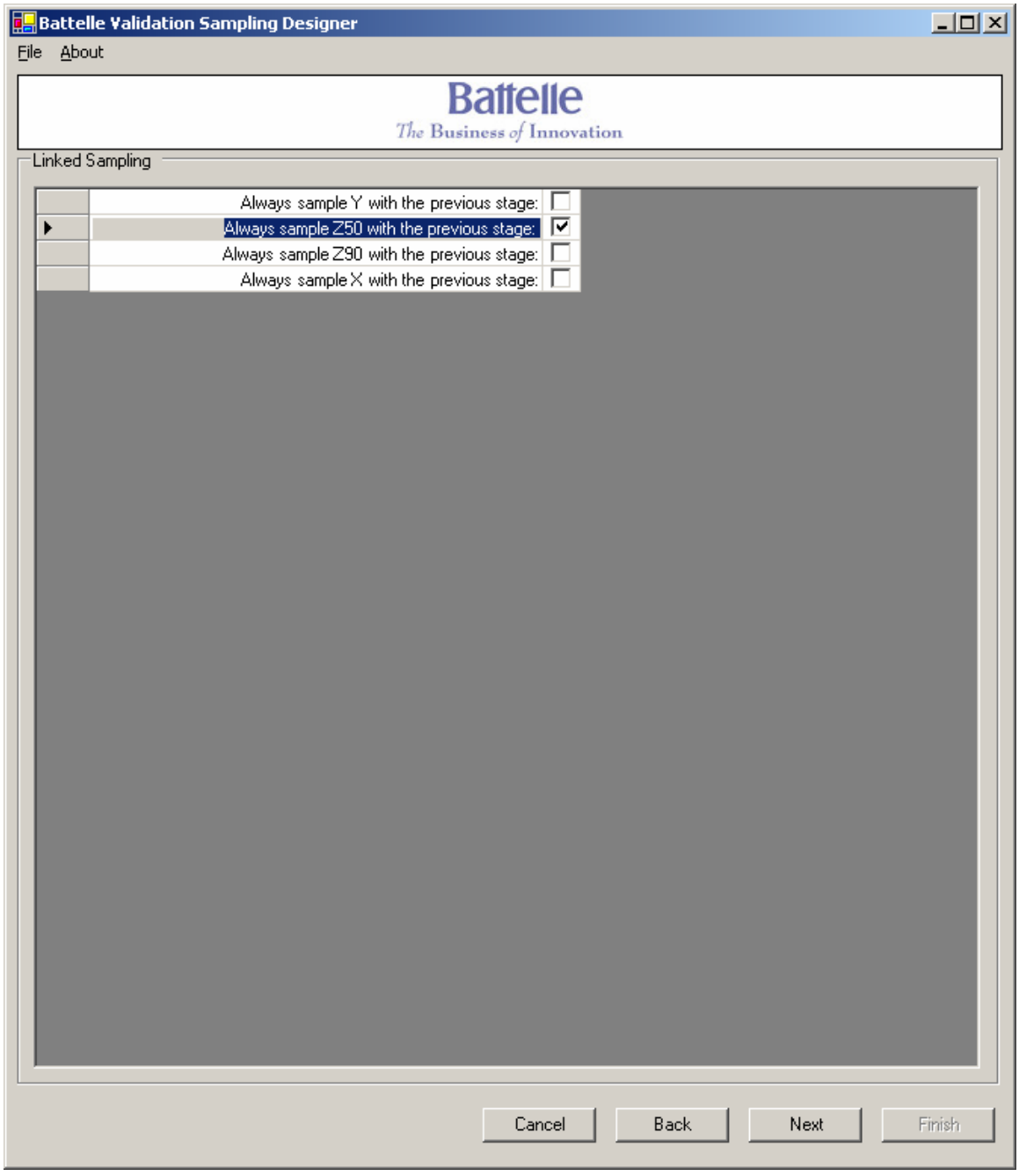


Figure 9 The Linked Stages of Sampling Constraint Window

5.8 Outcome and Covariate Dependent Sampling Design Input Window

The next step in the structured interview allows the user to choose covariate or outcome dependent sampling design options for variables at each stage within the design – subject to the constraints provided in earlier stages. As a continuation of the example initiated in Section 5.7, Figure 10 provides an example of how the user can provide input on covariate or outcome dependent sampling for a scenario which includes a health outcome (Y), two stages of surrogate measures (Z50 and Z90), and an exposure variable (X). In this example the design was constrained so that the surrogate exposure variable Z50 is measured on the identical subset of study participants that have the health outcome (Y) assessed.

To allow the user to select outcome and covariate dependent sampling design options, a triangular matrix of checkboxes is provided. The rows of the matrix represent each candidate variable in the design scenario which can be sampled as a function of other covariates or the outcome. The columns (with labels provided at the bottom of the matrix) represent candidate variables in the design for providing information for the sampling. Thus, the variable listed in the rows will be sampled as a function of all column variables whose boxes are checked.

In the example depicted in Figure 10, there is no row for the first surrogate exposure measure because of the constraint that Z50 is measured on the same subset of study participants as Y. The Stage 2 surrogate exposure measure (Z90) can be selected as a function of the outcome (Y) and/or the Stage 1 surrogate exposure measure (Z50). Since neither of these boxes in the first row is checked, the Stage 2 surrogate measure will be sampled at random. The exposure variable (X), which is measured in the final stage of sampling, can be selected as a function of the outcome (Y), the Stage 1 surrogate exposure measure (Z50), and/or the Stage 2 surrogate exposure measure (Z90). Since the box associated with the outcome (Y) in this last row is checked (and the other boxes in this row associated with the covariates Z50 and Z90 are unchecked), the design scenario will employ outcome dependent sampling for the exposure variable (X).

At the bottom of this window are three buttons: 'Next' which navigates the user to the next window in the structured interview, 'Back' which navigates the user to the previous window in the structured interview, and 'Cancel' which exits the user from the software. There is a fourth ('Finish') button at the bottom of the Window that will only become functional after the user has completed the structured interview process.

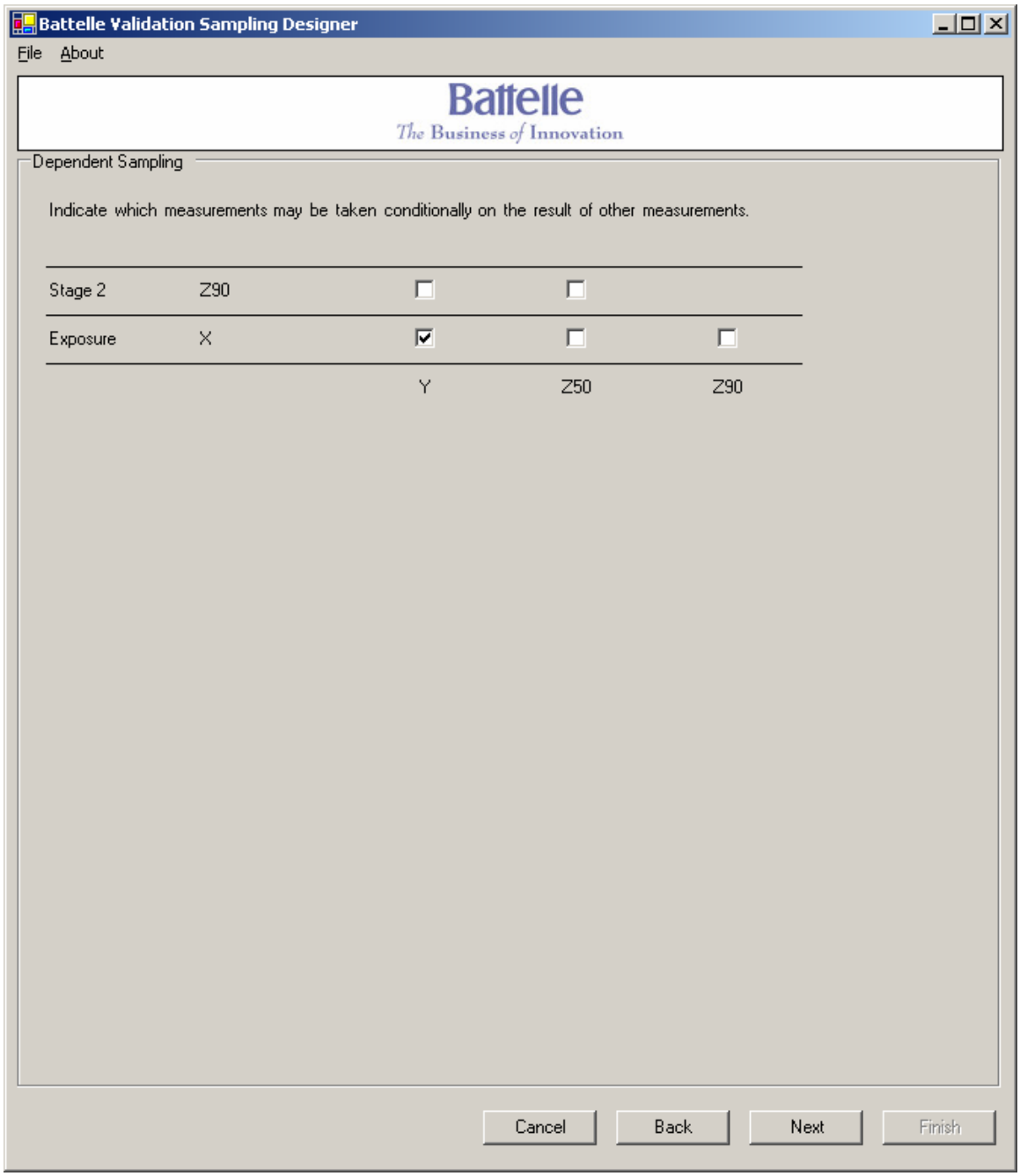


Figure 10 The Outcome and Covariate Dependent Sampling Design Input Window

Please note that the outcome/covariate dependent sampling matrix presented to the user in this window allows for dependent sampling designs at varying stages, and also multiple candidate surrogate measures of exposure to be explored within a single stage of the design. The outcome/covariate dependent sampling matrix will therefore be expanded to include all possible designs. As an example of this expansion, Figure 11 displays the outcome/covariate dependent sampling matrix for a design with two stages of surrogate sampling, an effect modifier (E), measurable exposure variable (X), two candidate surrogate measures in the first stage of surrogate sampling (Z30 and Z50), two candidate surrogate measures in the second stage of surrogate sampling (Z70 and Z90), and no constraints for linked sampling. In Figure 11, the Stage 2 variables (Z70 and Z90) can both be selected as a function of the Stage 1 variable Z50 – and the exposure variable can also be selected as a function of both the outcome (Y) and the Stage 1 surrogate (Z50). Table 1 provides a summary of the staged logit validation sampling equations that will result from this example, which will yield four designs based on the combination of multiple candidate surrogate exposure measures in Stages 1 and 2.

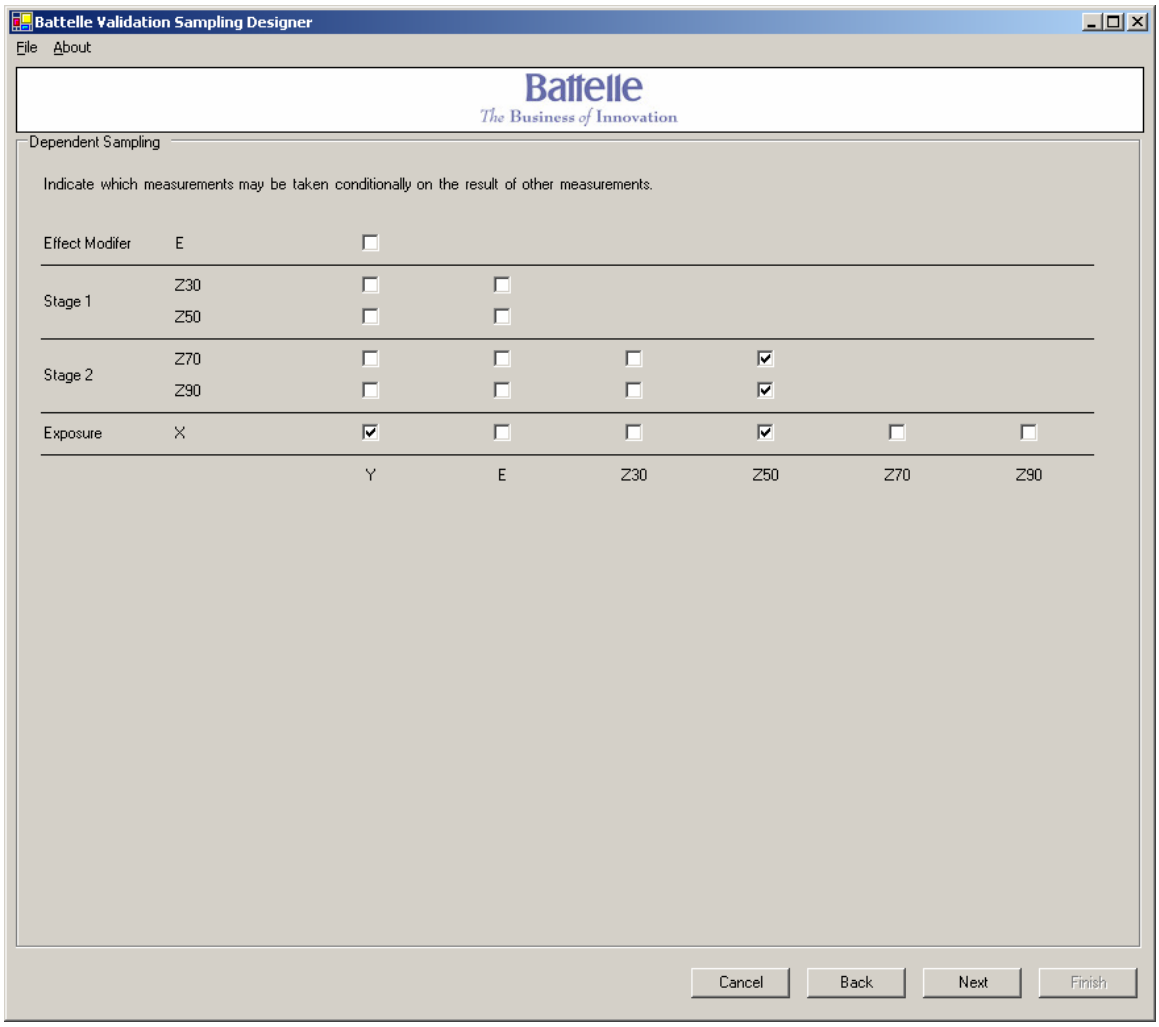


Figure 11 Example Outcome/Covariate Dependent Sampling Matrix Input for Example with 5 Stages and Multiple Candidate Surrogates

Table 1 Example Logit Validation Sampling Equations Corresponding to the Outcome/Covariate Dependent Sampling Designs Displayed in Figure 11

Stage 1 Variable	Target Measure	Stage 2 Variable	
		Z70	Z90
Z30	Y	$\text{logit}(\gamma_0) = \alpha_{00}$	$\text{logit}(\gamma_0) = \alpha_{00}$
	E	$\text{logit}(\gamma_1) = \alpha_{10}$	$\text{logit}(\gamma_1) = \alpha_{10}$
	Z30	$\text{logit}(\gamma_2) = \alpha_{20}$	$\text{logit}(\gamma_2) = \alpha_{20}$
	Z70/Z90	$\text{logit}(\gamma_3) = \alpha_{30}$	$\text{logit}(\gamma_3) = \alpha_{30}$
	X	$\text{logit}(\gamma_4) = \alpha_{40} + \alpha_{41} \cdot Y$	$\text{logit}(\gamma_4) = \alpha_{40} + \alpha_{41} \cdot Y$
Z50	Y	$\text{logit}(\gamma_0) = \alpha_{00}$	$\text{logit}(\gamma_0) = \alpha_{00}$
	E	$\text{logit}(\gamma_1) = \alpha_{10}$	$\text{logit}(\gamma_1) = \alpha_{10}$
	Z50	$\text{logit}(\gamma_2) = \alpha_{20}$	$\text{logit}(\gamma_2) = \alpha_{20}$
	Z70/Z90	$\text{logit}(\gamma_3) = \alpha_{30} + \alpha_{32} \cdot Z50$	$\text{logit}(\gamma_3) = \alpha_{30} + \alpha_{32} \cdot Z50$
	X	$\text{logit}(\gamma_4) = \alpha_{40} + \alpha_{41} \cdot Y + \alpha_{42} \cdot Z50$	$\text{logit}(\gamma_4) = \alpha_{40} + \alpha_{41} \cdot Y + \alpha_{42} \cdot Z50$

5.9 The Optimization Goals and Options Window

The Optimization Goals and Options Window, as seen in Figure 12, collects the last remaining input from the user prior to submitting the design characteristics to the analytical optimizer.

The first section of this Window provides a gateway for the user to input one or more optimization goals for the validation sampling design, using the same Add, Edit, or Delete options that appeared previously on the Inputting Surrogate Measures Window described in Section 5.6. The software allows for two general types of constrained optimization: (1) the user can specify a target total dollar budget to be spent on the data collection effort, and the constrained optimization will determine the set of α's that result in a design with the lowest possible standard error for the parameter of interest, or (2) the user can specify a target standard error for the parameter of interest, and the constrained optimization will determine the set of α's that result in a design with the lowest possible total dollar budget to be spent on the data collection effort. For this second type of optimization goal, the standard error is calculated as a function of the desired statistical power and size of the study (as well as a choice between a one-sided or two-sided test of the hypothesis), as seen in Figure 13.

Although the user can specify multiple optimization goals (which will lead to separate designs generated by the software), it is recommended to select only one type of optimization goal (constrain either budget or the standard error) in a single design submission because of anticipated differences in the Lagrange Multiplier between goal types (as discussed below).

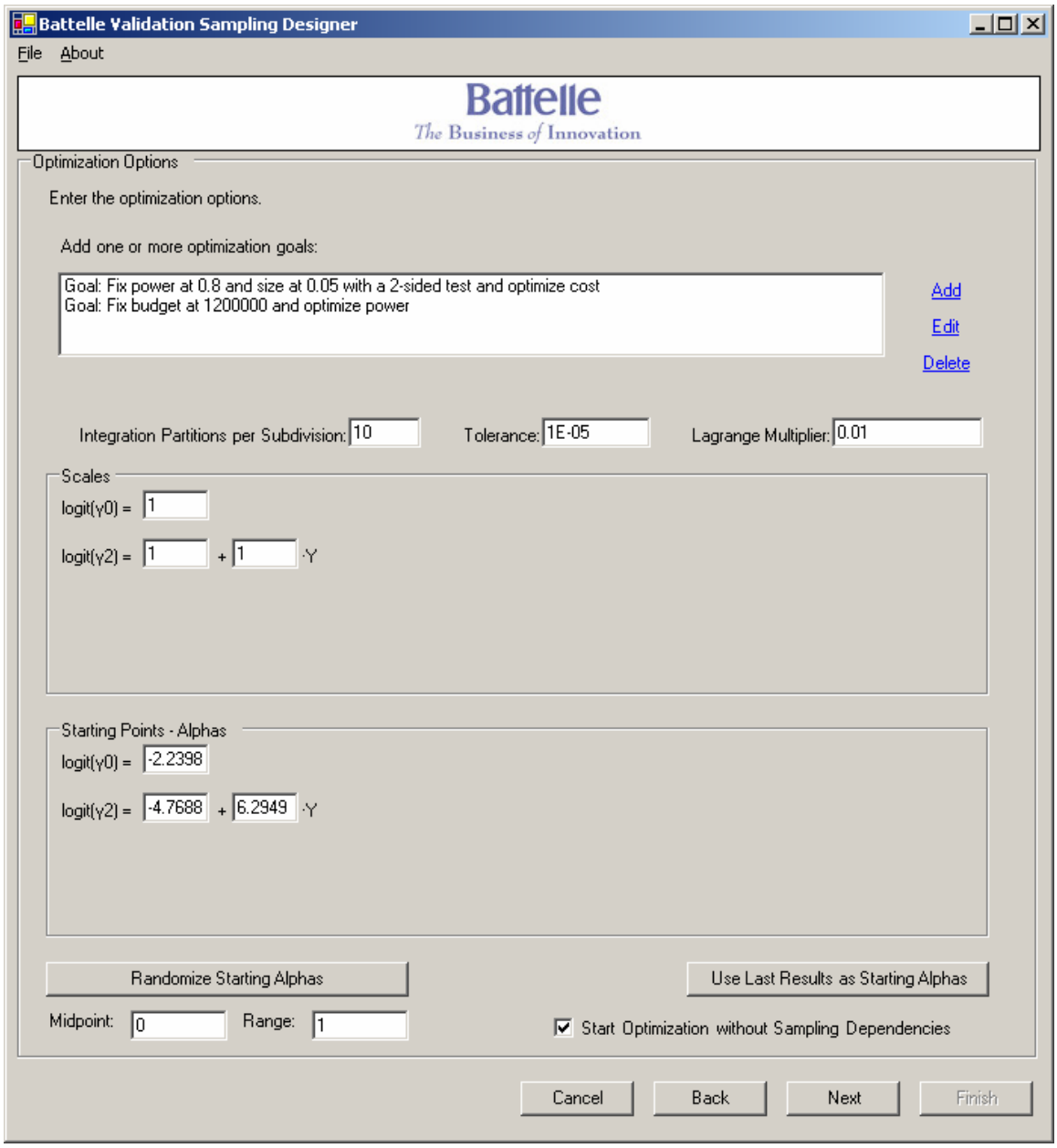


Figure 12 The Optimization Goals and Options Window

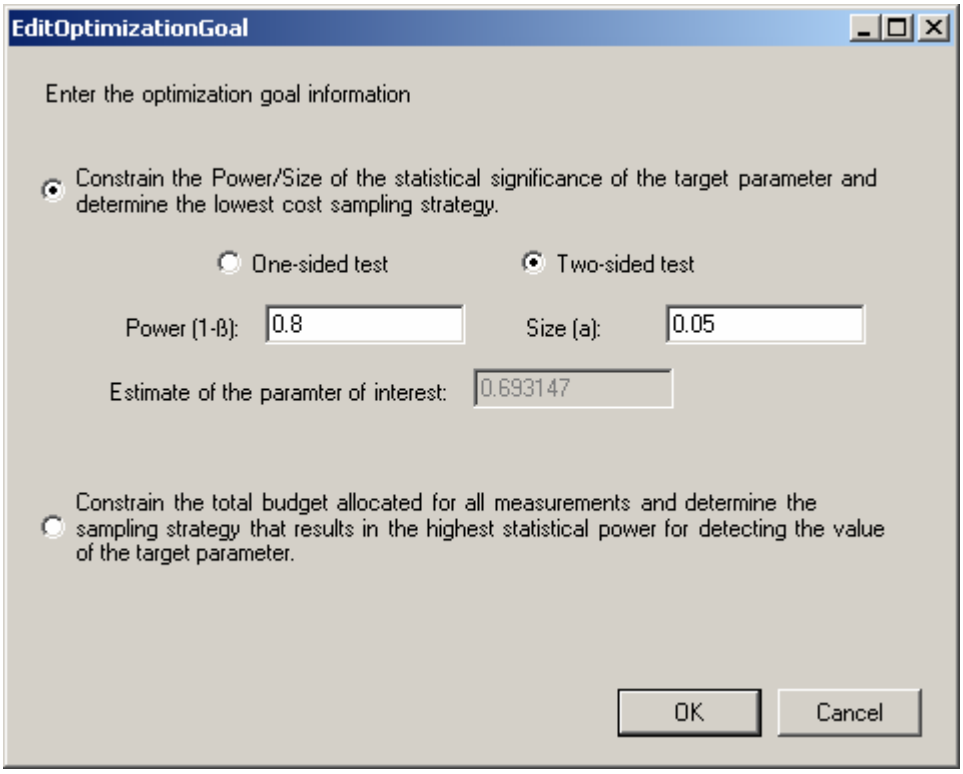


Figure 13 The Optimization Goal that Constrains the Standard Error of the Parameter of Interest can be Determined by Inputting the Desired Statistical Size, Power and Type (1-sided vs. 2-sided) of Test.

In the bottom of the first section of the Optimization Goals and Options Window are three optimization parameters as discussed below:

- ***Integration Partitions Per Dimension:*** This represents the number of partitions (subdivisions) in the univariate Gauss-Hermite integration (the number of segments in which the real axis is divided when numerical integration is conducted as part of the numeric optimization). A higher number of subdivisions leads to higher precision in the result. However, since the complexity of the algorithm is equal to the number of subdivisions to the power of the number of continuous variables, the optimization can become very slow as the number of subdivisions increases. Our initial research suggests a value between 10 and 12 should be used as the default value for initial searches – and that this number can be increased as the user attempts to optimize the design. To check if the result is sensitive to the number of partitions, the user should increase the number of partitions by 1 unit at a time to determine how sensitive the result is to this parameter.

- ***Tolerance:*** The tolerance represents the minimum percent decrease of the function before the searching algorithm stops. That is, if the percent reduction of

the function in the new point is less than the tolerance, then the iterative numerical optimization algorithm stops.

- ***Lagrange Multiplier:*** The numerical optimization is based on finding the minimum of the following formula:

$$L(\alpha) = f(\alpha) + \lambda \cdot (g(\alpha) - g_0)^2, \text{ where}$$

f(α) represents the quantity (budget or σ_β) that we want to minimize as a function of the sampling design parameters (α), g(α) represents the quantity (budget or σ_β) that we would like to constrain to a value of g_0 in the optimization, and λ represents the Lagrange multiplier (a multiplicative constant chosen to impose the constraint). The proper way to use the optimization tool is to start with a low value of λ and run a minimum search that loosely enforces the constraint, then increasing that value and run another minimum search using the result of the previous search as the starting point of the new one. The schematic algorithm could be,

1) Initialize the optimization by identifying a starting point for the design parameters (α_0) using a small value of the Lagrange multiplier (λ_0) which yields a solution that does not strictly enforce the budget or variance constraint (i.e. the constraint is achieved to within +/- 20 percent of the target value);
2) For iteration i (where i=1,…,n) – use the starting point for the design parameters from the previous iteration (α_{i-1}) and Lagrange multiplier (λ_i) (where $\lambda_i = c \cdot \lambda_{i-1}$, and c>1 (e.g. c=2)), identify the design parameters (α_i) that minimizes L

It is easy to check that if the above algorithm converges, it converges to a value α such that f(α) = min{f} and g(α) – g0 = 0. In practice, if we want to quickly check a sampling design strategy, most of the times one or two iterations of the above algorithm is enough when c=2. In that case lambda should be chosen as the lowest value such that the constraint is met within an acceptable error.

The middle sections of the Optimization Goals and Options Window allow the user to input values for the scales and alpha (design) parameters. These parameters can only be set by the user if the input to the design software results in a single optimization (i.e. the user cannot set these parameters in the event that multiple optimization goals or multiple candidate surrogate measures were specified within a single stage of sampling).

The scale parameters represent a vector of values representing the typical "length-scale" of each dimension. For example, the standard deviation can be considered the typical "length-scale" of a normal variable. The scale values need not be very precise, and in most cases the default value of 1 will suffice.

Towards the bottom of the Optimization Goals and Options Window are options that allow the user to (1) generate random values for the design parameter (α) at which to start the optimization search strategy, (2) utilize the design parameters from the previous optimization as starting values for the optimization, and (3) start the optimization without sampling dependencies. The advantages to utilizing these three options are described below:

- ***Advantages / Recommendations for picking random alpha's:*** If we pick many random starting points we have more chances to scan the whole domain and therefore to find the absolute minimum. This method might require a very high number of searches, and so it can be slow. Starting point randomization can also be useful to check the goodness of the result of a search. If that result is the absolute minimum, then a random starting point can at the most match it.

- ***Advantages to using last results as starting alpha's:*** This should be used either to refine the result of a previous search, or to start a new search with an increase Lagrange multiplier value. See "Lagrange Multiplier"

- ***Advantages to starting optimization without sampling dependencies:*** The alpha-space without sampling dependencies has fewer dimensions, and so the minimum search is quicker and more stable. If we suppose that the minimum without sampling dependencies can be used has a proxy to the minimum with the sampling dependency, than we can use the result of the minimum search without sampling dependencies as the starting point for the search with sampling dependency and ensure that the solution with sampling dependencies is at least as efficient as the solution without sampling dependencies.

Finally, at the bottom of this window are three buttons: 'Next' which navigates the user to the next window in the structured interview, 'Back' which navigates the user to the previous window in the structured interview, and 'Cancel' which exits the user from the software. There is a fourth ('Finish') button at the bottom of the Window that will only become functional after the user has completed the structured interview process.

5.10 The Input Review Window

Prior to submitting the design specifications to the analytic optimizer, a window will appear that will allow the user to review the design specifications that were input through the structured interview. There is a scroll bar to the right of the text box that allows the user to navigate through all of the design input. At the bottom of this window are three active buttons: 'Back' which navigates the user to the previous window in the structured interview, 'Cancel' which exits the user from the software, and 'Finish' which will close the structured interview and submit the design specifications to the analytical optimizer. There is a fourth ('Next') button at the bottom of the Window that is disabled.

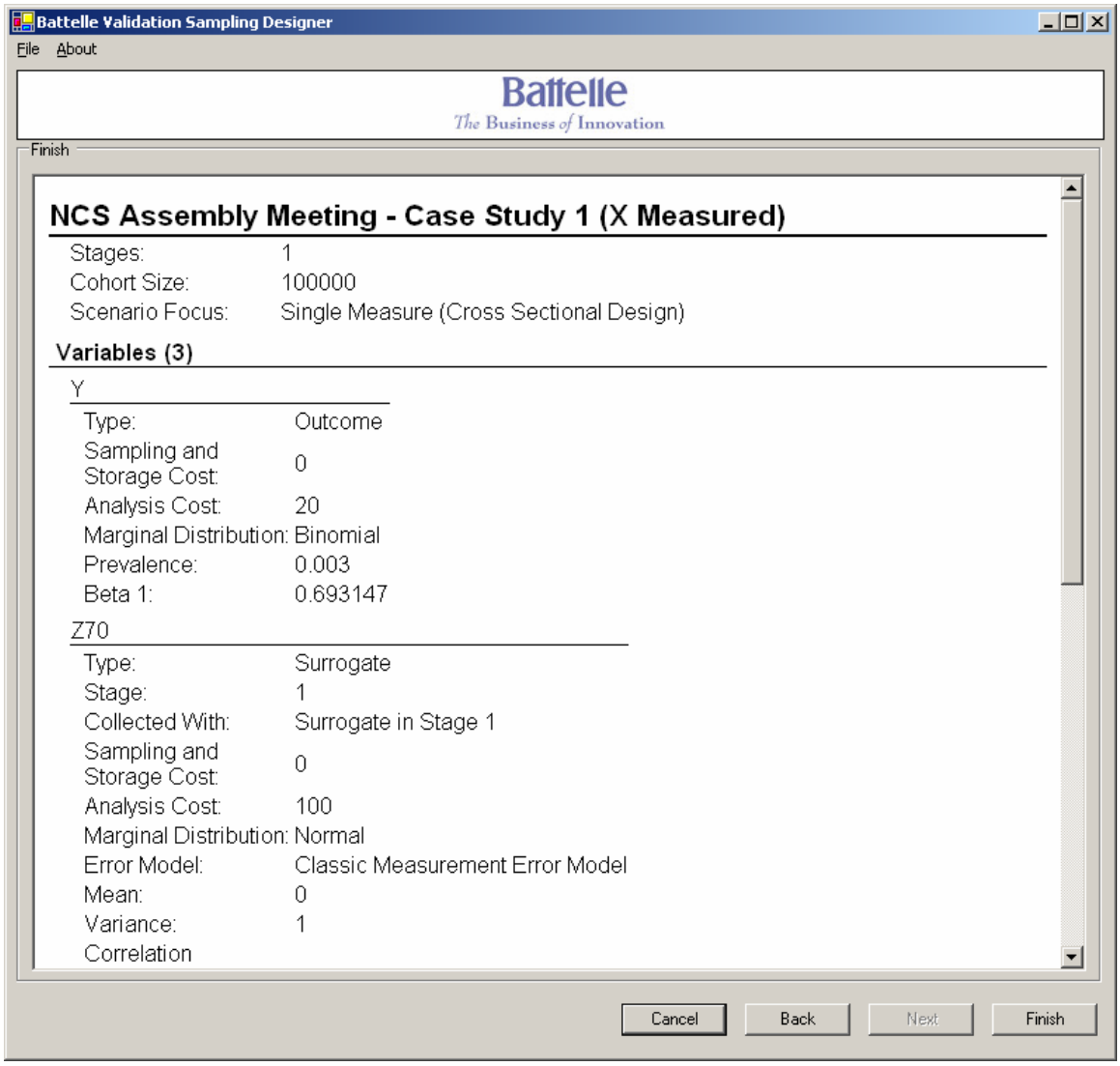
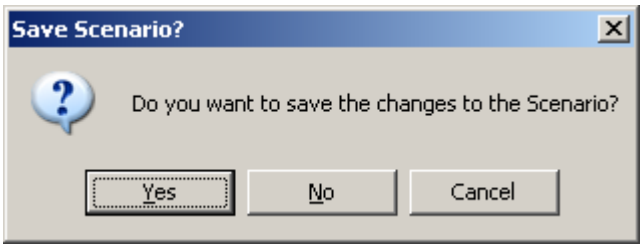


Figure 14 The Input Review Window allows the User to Review the Design Specifications prior to Submission to the Optimizer

When the user selects the 'Finish' button, a window will appear that will allow the user to save the current scenario to a file.



If the user was working on an existing file, selecting the 'Yes' button will automatically update the previously saved file. If the user was working on new design specifications, selecting the 'Yes' button will open a standard windows folder that allows the user to provide a name and location for the saved design input. Selecting the 'No' button will allow the design input to be submitted to the analyzer without saving the input to a file. Selecting the 'Cancel' button will allow the user to go back to the Review Input Window without submitting the design specifications to the optimizer.

5.11 Scenario Running Window

Once the design input is submitted to the analyzer, the Scenario Running Window will appear. This window includes the following six subwindows:

1. ***Current Processing Progress:*** This window provides a display of optimization progress as the analyzer is engaged. Each row of data in this window corresponds to one of the iterative steps of the numeric optimization. The first number displayed in each row corresponds to the design scenario being optimized (starting with a value of zero), the second number represents the iteration step, the third number represents the value of the function (see the Section 5.9 discussion on the Lagrange Multiplier) being optimized at the starting point in the iteration step, and the fourth number represents the value of the function at the stopping point in the iteration step. When the absolute value of the difference between the third and fourth numbers in the display is less than the specified tolerance (see Section 5.9), the optimizer will have satisfied the convergence criteria and the results will be displayed in the subsequent windows.

2. ***Variables Included:*** This window provides a listing of all the variables included in the design scenario being optimized. For example, in Figure 14 there are four candidate surrogate measures being screened (Z30, Z50, Z70, and Z90) – and each design considered includes the health outcome (Y), the exposure variable (X), and one of the four candidate surrogate measures.

3. ***Expected Values:*** This window provides the expected value of the total budget and standard error for the parameter of interest associated with each design. In the example displayed in Figure 14 – the standard error of the parameter of

interest was constrained, allowing the optimizer to determine the lowest cost validation sampling design associated with each candidate surrogate measure. If screening multiple candidate designs (as is the case displayed in Figure 14), the user can sort the resulting designs by either 'Total Cost' or 'Standard Error' by clicking on either column heading in this window. The output displayed in all other windows will be adjusted to the appropriate order based on this re-ranking.

4. ***Sampling Probabilities:*** This provides the average sampling probabilities associated with measures at each stage of sampling. If the staged sampling probability is constrained to the same value as a previous stage, the sampling probability associated with that previous stage is displayed. In the example displayed in Figure 14, the staged probability for each surrogate measure was constrained to the same fraction of the cohort that is sampled for the health outcome (Y). Therefore, the sampling probabilities for the health outcome Y and the appropriate surrogate measure (Z30, Z50, Z70, or Z90) are the same within each row. If screening multiple candidate designs (as is the case displayed in Figure 14), the user can sort the resulting designs by the sampling probabilities associated with each candidate variable by clicking on the variable name column headings in this window. The output displayed in all other windows will be adjusted to the appropriate order based on this re-ranking.

5. ***Expected Sample Sizes:*** This provides the expected number of study subjects anticipated to be sampled for each measure within a design. This number is simply the rounded product between the Sampling Probabilities displayed in the previous window and the size of the cohort defined in the High Level Design Summary Window (see Section 5.2). If screening multiple candidate designs (as is the case displayed in Figure 14), the user can sort the resulting designs by the expected sample sizes associated with each candidate variable by clicking on the variable name column headings in this window. The output displayed in all other windows will be adjusted to the appropriate order based on this re-ranking.

6. ***Sampling Equations:*** This window provides the logistic regression sampling equations associated with each design (see Section 4 and 5.8 for a description). In the example depicted in Figure 14, the logistic regression sampling equations include a simple intercept for the first stage of sampling for the health outcome and surrogate measure, and both outcome and covariate dependent sampling are employed at the second stage of sampling.

At the bottom of this window are two buttons: 'Save Results' - which saves the validation sampling design output into a separate file, and 'Close' – which simply closes the Scenario Running Window and returns the user back to the first screen in the structured interview (the High Level Design Summary Window, as described in Section 5.2). Appendix A provides an example of the output produced by the Battelle Validation Sampling Design Software Tool.

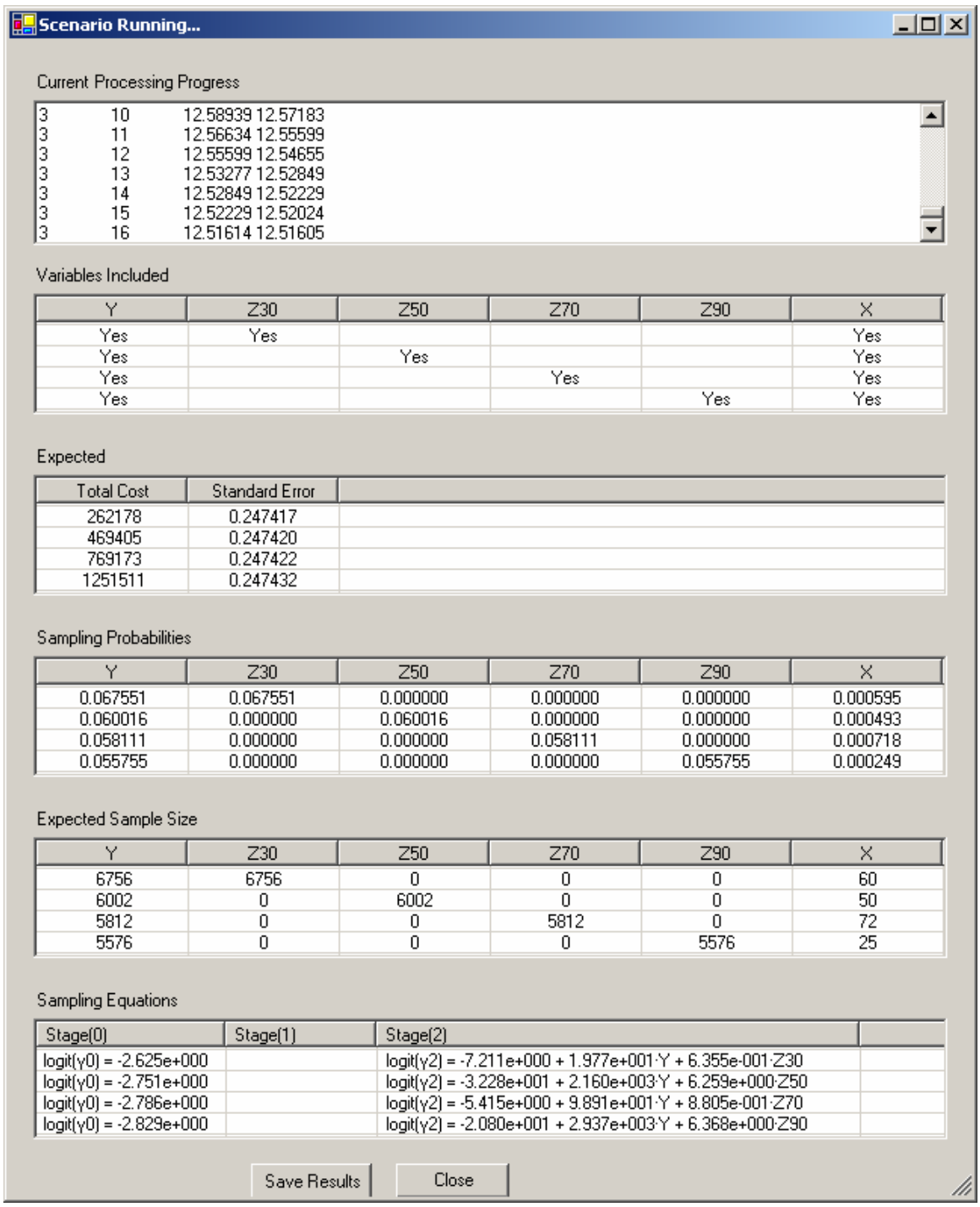


Figure 15 Example of the Scenario Running Window

Appendix A

Example Output from the Battelle Validation Sampling Design Software

NCS Assembly Meeting - Case Study 1 (X Measured)

Stages:	1
Cohort Size:	100000
Scenario Focus:	Single Measure (Cross Sectional Design)

Variables (6)

Y

Type:	Outcome
Sampling and Storage Cost:	0
Analysis Cost:	20
Marginal Distribution:	Binomial
Prevalence:	0.003
Beta 1:	0.693147

Z30

Type:	Surrogate
Stage:	1
Collected With:	Surrogate in Stage 1
Sampling and Storage Cost:	0
Analysis Cost:	10
Marginal Distribution:	Normal
Error Model:	Classic Measurement Error Model
Mean:	0
Variance:	1
Correlation Coefficient With Exposure:	0.3

Z50

Type:	Surrogate
Stage:	1
Collected With:	Surrogate in Stage 1
Sampling and Storage Cost:	0
Analysis Cost:	50
Marginal Distribution:	Normal
Error Model:	Classic Measurement Error Model

Mean:	0
Variance:	1
Correlation Coefficient With Exposure:	0.5

Z70

Type:	Surrogate
Stage:	1
Collected With:	Surrogate in Stage 1
Sampling and Storage Cost:	0
Analysis Cost:	100
Marginal Distribution:	Normal
Error Model:	Classic Measurement Error Model
Mean:	0
Variance:	1
Correlation Coefficient With Exposure:	0.7

Z90

Type:	Surrogate
Stage:	1
Collected With:	Surrogate in Stage 1
Sampling and Storage Cost:	0
Analysis Cost:	200
Marginal Distribution:	Normal
Error Model:	Classic Measurement Error Model
Mean:	0
Variance:	1
Correlation Coefficient With Exposure:	0.9

X

Type:	Exposure
Collected With:	Exposure
Sampling and Storage Cost:	0
Analysis Cost:	1000
Marginal Distribution:	Normal
Mean:	0
Variance:	1

Optimization Goals (1)

Goal: Fix power at 0.8 and size at 0.05 with a 2-sided test and optimize cost

Sampling Dependencies

Z30:	None
Z50:	None
Z70:	None
Z90:	None
X:	Y, Z70, Z30, Z50, Z90

Optimization Options

Subdivisions:	10
Tolerance:	1E-05
Lagrange Multiplier:	0.001
Gammas:	False, True, True, True, True, False

Scenario Results (4)

Input 1

Optimization Parameters

Constraint Type:	Variance of parameter estimate is constrained - Determine lowest cost sampling plan
Target Value for Constrained Optimization:	Standard Deviation of parameter estimate constrained to 0.247412
Lagrange Multiplier:	0.001
Tolerence:	0.00001
Start With Sampling Dependencies:	true
Starting Alphas:	$\text{logit}(\gamma_0) = 0$ $\text{logit}(\gamma_2) = 0 + 0\cdot + 0\cdot$
Scales:	$\text{logit}(\gamma_0) = 1$ $\text{logit}(\gamma_2) = 1 + 1\cdot + 1\cdot$

Sampling Design Parameters

Integration Subdivisions: 10

Y

Type:	Outcome
Stage:	0
Costs:	Cost Associated with Sampling of Y is $0 Cost Associated with Analysis of Y is $20 Total Cost at Stage 0 is $20
Distribution Type:	Binomial
Parameters:	Y = -6.044522 + 0.693147·X

Z30

Type:	Surrogate
Stage:	1
Costs:	Cost Associated with Sampling of Z30 is $0 Cost Associated with Sampling of Other Measures at Stage 1 is $0 Cost Associated with Analysis of Z30 is $10 Total Cost at Stage 1 is $10
Distribution Type:	Normal
Measurement Error Model:	true
Parameters:	Z30 = X + Error; σ^2_{Error} = 10.111111

X

Type:	Exposure
Stage:	2
Costs:	Cost Associated with Sampling of X is $0 Cost Associated with Analysis of X is $1000 Total Cost at Stage 2 is $1000
Distribution Type:	Normal
Parameters:	X ~ N(0, 1)

Sampling Dependencies

X:	Sample Dependent on Y, Z30
Gamma Constraints:	Y and Z30 will be sampled together

Output

Standard Deviation of B_1:	0.247417
Total Cost of Design:	$262,178
Sample Sizes:	Stage 0: Y sampled for 6755 subjects Stage 1: Z30 sampled for 6755 subjects Stage 2: X sampled for 60 subjects
Sampling Equations:	$\text{logit}(\gamma_{0,1}) = -2.62493$ $\text{logit}(\gamma_2) = -7.21107 + 19.7679 \cdot Y + 0.635458 \cdot Z30$

Input 2

Optimization Parameters

Constraint Type:	Variance of parameter estimate is constrained - Determine lowest cost sampling plan
Target Value for Constrained Optimization:	Standard Deviation of parameter estimate constrained to 0.247412
Lagrange Multiplier:	0.001
Tolerence:	0.00001
Start With Sampling	true

Dependencies:

Starting Alphas:	$\text{logit}(\gamma_0) = 0$ $\text{logit}(\gamma_2) = 0 + 0\cdot + 0\cdot$
Scales:	$\text{logit}(\gamma_0) = 1$ $\text{logit}(\gamma_2) = 1 + 1\cdot + 1\cdot$

Sampling Design Parameters

Integration Subdivisions: 10

Y

Type:	Outcome
Stage:	0
Costs:	Cost Associated with Sampling of Y is $0 Cost Associated with Analysis of Y is $20 Total Cost at Stage 0 is $20
Distribution Type:	Binomial
Parameters:	Y = -6.044522 + 0.693147·X

Z50

Type:	Surrogate
Stage:	1
Costs:	Cost Associated with Sampling of Z50 is $0 Cost Associated with Sampling of Other Measures at Stage 1 is $0 Cost Associated with Analysis of Z50 is $50 Total Cost at Stage 1 is $50
Distribution Type:	Normal
Measurement Error Model:	true
Parameters:	Z50 = X + Error; $\sigma^2_{Error} = 3$

X

Type:	Exposure
Stage:	2
Costs:	Cost Associated with Sampling of X is $0 Cost Associated with Analysis of X is $1000 Total Cost at Stage 2 is $1000
Distribution Type:	Normal
Parameters:	X ~ N(0, 1)

Sampling Dependencies

X:	Sample Dependent on Y, Z50
Gamma Constraints:	Y and Z50 will be sampled together

Output

Standard Deviation of 0.24742

B_1:	
Total Cost of Design:	$469,405
Sample Sizes:	Stage 0: Y sampled for 6002 subjects Stage 1: Z50 sampled for 6002 subjects Stage 2: X sampled for 49 subjects
Sampling Equations:	$\text{logit}(\gamma_{0,1}) = -2.75125$ $\text{logit}(\gamma_2) = -32.2786 + 2160.07 \cdot Y + 6.25869 \cdot Z50$

Input 3

Optimization Parameters

Constraint Type:	Variance of parameter estimate is constrained - Determine lowest cost sampling plan
Target Value for Constrained Optimization:	Standard Deviation of parameter estimate constrained to 0.247412
Lagrange Multiplier:	0.001
Tolerence:	0.00001
Start With Sampling Dependencies:	true
Starting Alphas:	$\text{logit}(\gamma_0) = 0$ $\text{logit}(\gamma_2) = 0 + 0 \cdot + 0 \cdot$
Scales:	$\text{logit}(\gamma_0) = 1$ $\text{logit}(\gamma_2) = 1 + 1 \cdot + 1 \cdot$

Sampling Design Parameters

Integration Subdivisions: 10

Y

Type:	Outcome
Stage:	0
Costs:	Cost Associated with Sampling of Y is $0 Cost Associated with Analysis of Y is $20 Total Cost at Stage 0 is $20
Distribution Type:	Binomial
Parameters:	$Y = -6.044522 + 0.693147 \cdot X$

Z70

Type:	Surrogate
Stage:	1
Costs:	Cost Associated with Sampling of Z70 is $0 Cost Associated with Sampling of Other Measures at Stage 1 is $0 Cost Associated with Analysis of Z70 is $100 Total Cost at Stage 1 is $100
Distribution Type:	Normal

Measurement Error Model:	true
Parameters:	Z70 = X + Error; σ^2_{Error} = 1.040816

X

Type:	Exposure
Stage:	2
Costs:	Cost Associated with Sampling of X is $0 Cost Associated with Analysis of X is $1000 Total Cost at Stage 2 is $1000
Distribution Type:	Normal
Parameters:	X ~ N(0, 1)

Sampling Dependencies

Z70:	Sampled at Random
X:	Sample Dependent on Y, Z70
Gamma Constraints:	Y and Z70 will be sampled together

Output

Standard Deviation of B_1:	0.247422
Total Cost of Design:	$769,173
Sample Sizes:	Stage 0: Y sampled for 5811 subjects Stage 1: Z70 sampled for 5811 subjects Stage 2: X sampled for 72 subjects
Sampling Equations:	$\text{logit}(\gamma_{0,1}) = -2.78553$ $\text{logit}(\gamma_2) = -5.4152 + 98.9084 \cdot Y + 0.880526 \cdot Z70$

Input 4

Optimization Parameters

Constraint Type:	Variance of parameter estimate is constrained - Determine lowest cost sampling plan
Target Value for Constrained Optimization:	Standard Deviation of parameter estimate constrained to 0.247412
Lagrange Multiplier:	0.001
Tolerence:	0.00001
Start With Sampling Dependencies:	true
Starting Alphas:	$\text{logit}(\gamma_0) = 0$ $\text{logit}(\gamma_2) = 0 + 0 \cdot + 0 \cdot$
Scales:	$\text{logit}(\gamma_0) = 1$ $\text{logit}(\gamma_2) = 1 + 1 \cdot + 1 \cdot$

Sampling Design Parameters

Integration Subdivisions: 10

Y

Type:	Outcome
Stage:	0
Costs:	Cost Associated with Sampling of Y is $0 Cost Associated with Analysis of Y is $20 Total Cost at Stage 0 is $20
Distribution Type:	Binomial
Parameters:	Y = -6.044522 + 0.693147·X

Z90

Type:	Surrogate
Stage:	1
Costs:	Cost Associated with Sampling of Z90 is $0 Cost Associated with Sampling of Other Measures at Stage 1 is $0 Cost Associated with Analysis of Z90 is $200 Total Cost at Stage 1 is $200
Distribution Type:	Normal
Measurement Error Model:	true
Parameters:	Z90 = X + Error; $\sigma^2_{Error} = 0.234568$

X

Type:	Exposure
Stage:	2
Costs:	Cost Associated with Sampling of X is $0 Cost Associated with Analysis of X is $1000 Total Cost at Stage 2 is $1000
Distribution Type:	Normal
Parameters:	X ~ N(0, 1)

Sampling Dependencies

X:	Sample Dependent on Y, Z90
Gamma Constraints:	Y and Z90 will be sampled together

Output

Standard Deviation of B_1:	0.247432
Total Cost of Design:	$1,251,510
Sample Sizes:	Stage 0: Y sampled for 5576 subjects Stage 1: Z90 sampled for 5576 subjects Stage 2: X sampled for 25 subjects
Sampling Equations:	$\text{logit}(\gamma_{0,1}) = -2.82941$ $\text{logit}(\gamma_2) = -20.8029 + 2937.5 \cdot Y + 6.3685 \cdot Z90$